Perfect Postrior Inference for Mixture Models

Kasper K. Berthelsen

Department of Mathematical Scieces
Aalborg University

Joint work with Gareth O. Roberts and Laird A. Breyer.

Warwick, March 2009

Overview

- The mixture problem
- Random update function
- Perfect simulation
- Bounding sets
- Simulation results

Mixture Problem

Data distribution: r component mixture

$$\eta_1, \ldots, \eta_n \overset{IID}{\sim} \sum_{k=1}^{r} m_k \pi(\cdot|k)$$

where $\pi(\cdot|1), \ldots, \pi(\cdot|r)$ are r probability densities and $m_1, \ldots, m_r$ are unknown weights with $\sum_{k=1}^{r} m_k = 1$.

Data Augmentation (Diebolt & Robert, 1994)

Augment allocation variable $z_1, \ldots, z_n \in \{1, \ldots, r\}$ and consider the state vector

$$\mathbf{x} = (z_1, \ldots, z_n, m_1, \ldots, m_r).$$

Alternative formulation of data distribution:
Assume $z_1, \ldots, z_n$ are IID with

$$\mathbb{P}(z_s = k) = m_k$$

and $\eta_1, \ldots, \eta_n$ are conditional independent given $z = (z_1, \ldots, z_n)$ with

$$\eta_s | z_s = k \sim \pi(\cdot | k)$$

Bayesian Analysis

Likelihood:

$$\pi(\eta|\mathbf{x}) = \prod_{s=1}^{n} \pi(\eta_s|z_s)$$

Prior for $\mathbf{x} = (\mathbf{z}, \mathbf{m})$: We assume

$$\pi(\mathbf{m}) = \mathcal{D}(1, \ldots, 1)$$

where $\mathcal{D}(\alpha_1 + 1, \ldots, \alpha_r + 1)$ denotes the Dirichlet distribution with density

$$f_\alpha(\mathbf{x}) = \frac{\Gamma(r + \alpha_1 + \cdots + \alpha_r)}{\Gamma(1 + \alpha_1) \cdots \Gamma(1 + \alpha_r)} m_1^{\alpha_1} \cdots m_r^{\alpha_r}.$$

and conditional on $\mathbf{m}$

$$\pi(\mathbf{z}|\mathbf{m}) = \prod_{i=1}^{s} \mathbb{P}(z_s = k|\mathbf{m}), \text{ where } P(z_s = k|\mathbf{m}) = m_k.$$

Posterior

Posterior:

$$\pi(\mathbf{x}|\eta) \propto \prod_{s=1}^{n} \pi(\eta_s|z_s) \prod_{k=1}^{r} m_k^{N_k(x)}$$

where $N_k(x) = \#\{s : z_s = k\}$ is the number of points allocated to the kth component.

Sample $\pi(\mathbf{x}|\eta)$ using Gibbs sampling and the updating scheme

$$\mathbf{x} = (\mathbf{z}, \mathbf{m}) \to (\mathbf{z}, \mathbf{m}') \to \mathbf{x}' = (\mathbf{z}', \mathbf{m}')$$

Full conditionals:

- $\mathbf{m}'|\mathbf{z} \sim \mathcal{D}(N_1(\mathbf{x}) + 1, \ldots, N_r(\mathbf{x}) + 1)$
- $\mathbb{P}(z_s' = k|\mathbf{m}', z_s' \geq k) = \frac{m_k' \pi(\eta_s|k)}{\sum_{j=k}^{r} m_j' \pi(\eta_s|j)}$

Gibbs Sampler as Random Function

Construct random function F so that

$$\mathbf{x}' = (\mathbf{z}', \mathbf{m}') = F(\mathbf{x}) = (\mathbf{Z}(\mathbf{x}), \mathbf{M}(\mathbf{x})).$$

Notice that

- $\mathbf{M}(\mathbf{x})$ only depends on $\mathbf{x}$ through $N_1(\mathbf{x}), \ldots, N_r(\mathbf{x})$.
- $\mathbf{Z}(\mathbf{x})$ only depends on $\mathbf{x}$ through $\mathbf{M}(\mathbf{x})$.

Constructing $\mathbf{M}(\mathbf{x})$

We generate the random function $\mathbf{M}(\mathbf{x}) = (M_1(\mathbf{x}), \ldots, M_r(\mathbf{x}))$ as follows.

Generate IID random functions $G_i : \mathbb{N} \to \mathbb{R}$, $i = 1, \ldots, r$ so that

1. $G_i(k+1) \geq G_i(k)$ and
2. $G_i(k) \sim \Gamma(k)$.

Here $\Gamma(k)$ denote a gamma distribution with scale parameter 1 and shape parameter k.

Let

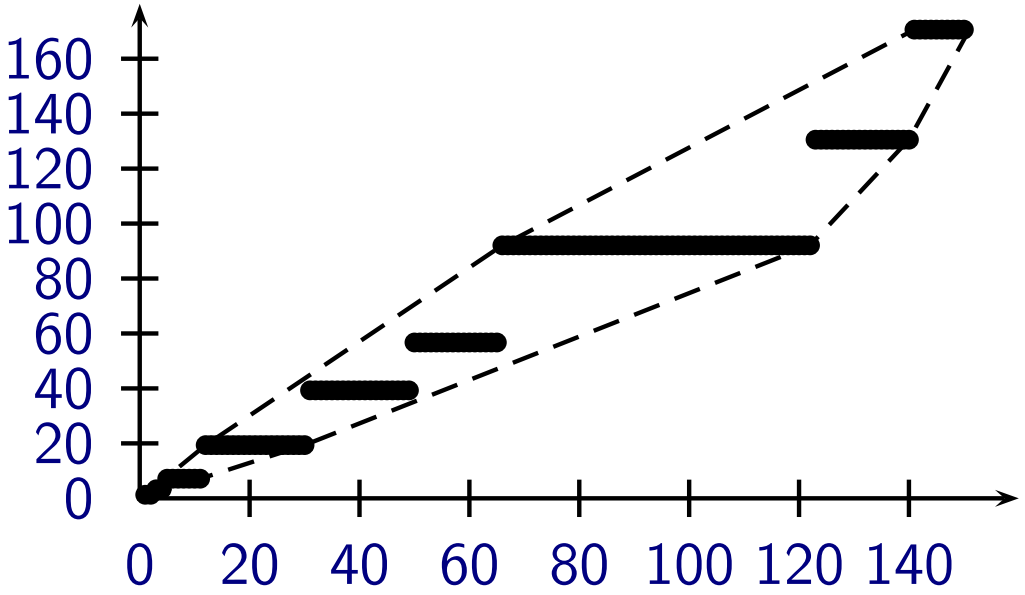
$$M_k(\mathbf{x}) = \frac{G_k(N_k(\mathbf{x}) + 1)}{\sum_{j=1}^{r} G_j(N_j(\mathbf{x}) + 1)}.$$

Then $\mathbf{M}(\mathbf{x}) \sim \mathcal{D}(N_1(\mathbf{x}) + 1, \ldots, N_r(\mathbf{x}) + 1)$.

Comments on G

In practise $G_i(k+1) = G_i(k)$ for many k.

Empirically it seems that $\#\{G(k) : k = 1, \ldots, n+1\}$ is $O(\sqrt{n})$.



Constructing $\mathbf{Z}(\mathbf{x})$

For each Z_s generate $\xi_{s,1}, \ldots, \xi_{s,r} \overset{IID}{\sim} U[0,1]$.

For each component $k = 1, \ldots, r$ in turn, we accept $Z_s(\mathbf{x}) = k$ if

- no other component $j < k$ has yet been accepted, and
- $\xi_{s,k} < \pi_k(\eta_s) M_k(\mathbf{x}) / \sum_{j=k}^{r} \pi_j(\eta_s) M_j(\mathbf{x})$.

Read Once Coupling From The Past (Wilson 2000)

Assume that $C : \Omega \to \Omega$ is a random function that preserves stationarity w.r.t. target distribution Π, ie.

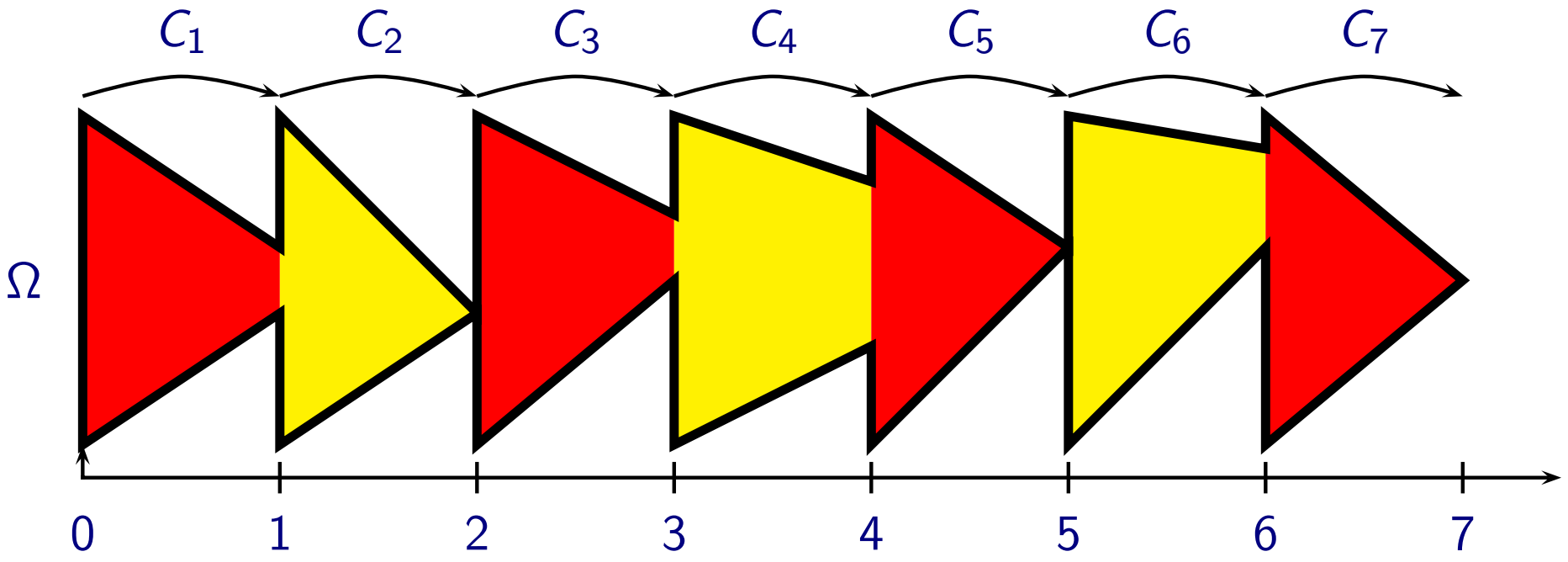
$$\int_{\Omega} \mathbb{P}(C(x) \in A)\Pi(dx) = \Pi(A).$$

Assume there is a positive probability of C being coalescent, ie. $\#C(\Omega) = 1$.

Here $C(\Omega) = \{C(x) : x \in \Omega\}$ is the image of C and $\#$ denotes cardinality.

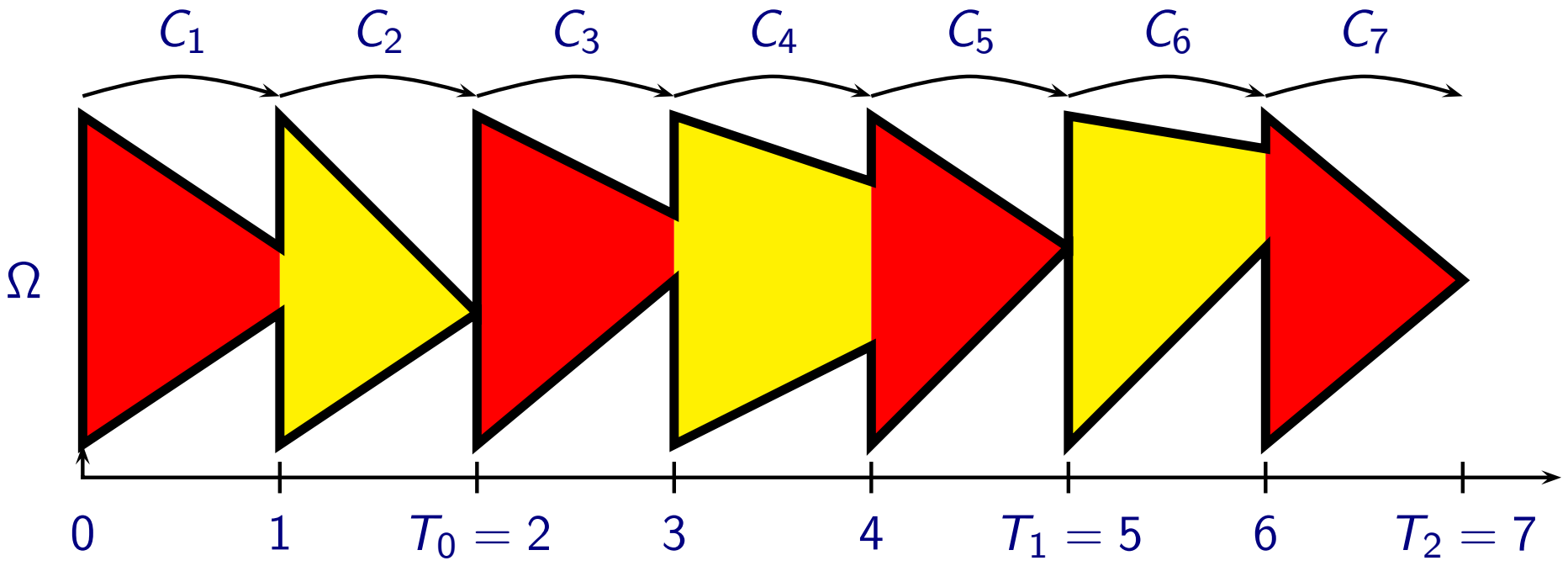
Cartoon CFTP

1. Generate independent realisations $C_1, C_2, \ldots$ of C
2. Let $T_0, T_1, \ldots$ denote indices of coalescent functions, ie. $C_{T_i}(\Omega)$ is coalescent.
3. For $i = 1, 2, \ldots$ let $x_i = C_{T_{i-1}} \circ \cdots \circ C_{T_{i-1}}(\Omega)$



Cartoon CFTP

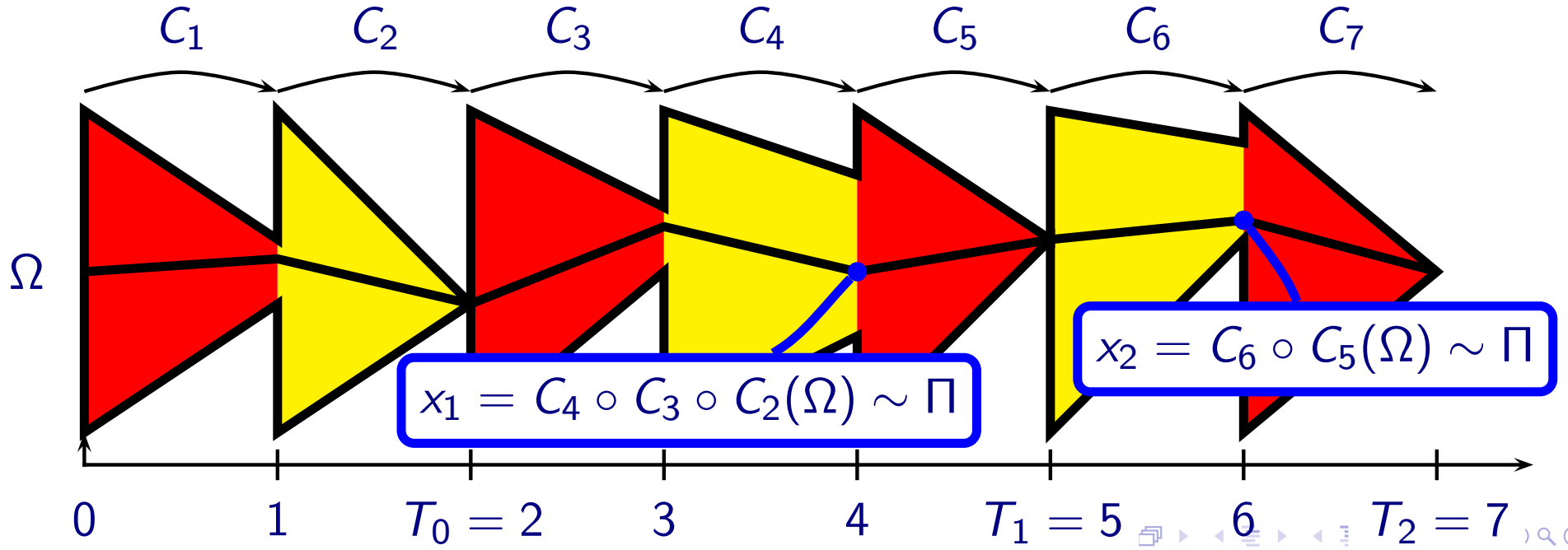
1. Generate independent realisations $C_1, C_2, \ldots$ of C
2. Let $T_0, T_1, \ldots$ denote indices of coalescent functions, ie. $C_{T_i}(\Omega)$ is coalescent.
3. For $i = 1, 2, \ldots$ let $x_i = C_{T_{i-1}} \circ \cdots \circ C_{T_{i-1}}(\Omega)$



Cartoon CFTP

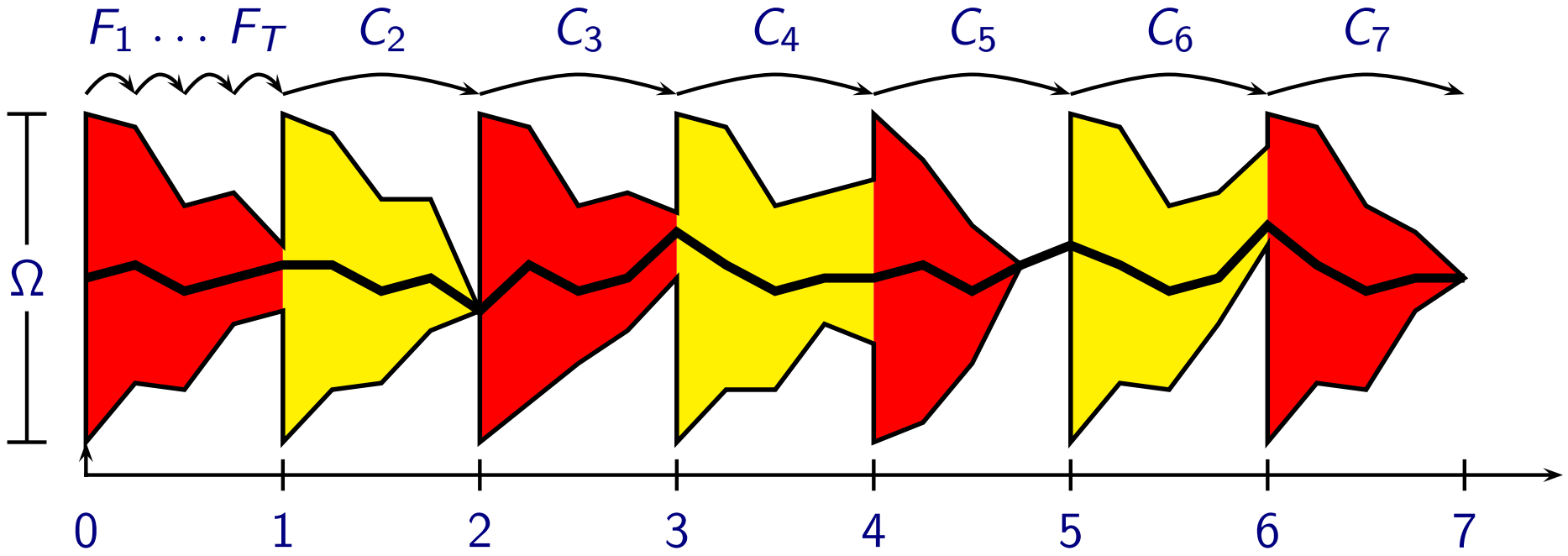
1. Generate independent realisations $C_1, C_2, \ldots$ of C
2. Let $T_0, T_1, \ldots$ denote indices of coalescent functions, ie. $C_{T_i}(\Omega)$ is coalescent.
3. For $i = 1, 2, \ldots$ let $x_i = C_{T_i - 1} \circ \cdots \circ C_{T_{i-1}}(\Omega)$

Then $x_1, x_2, \ldots$ are an IID sample from Π.



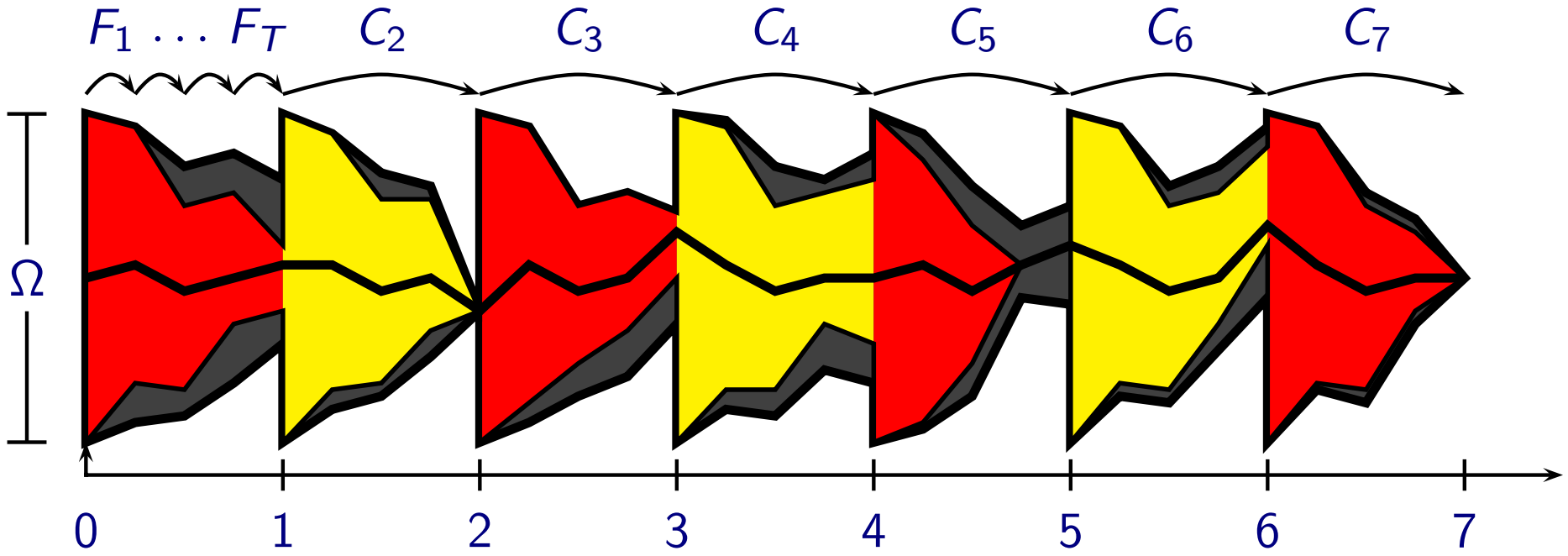
CFTP in Practice

- C is a compound update function $C = F_T \circ \cdots \circ F_1$
- Assume we can determine bounding set $W_t \supseteq C_t(\Omega)$.
- Redefine T_i, so W_{T_i} is "coalescent".
- Then $x_1, x_2, \ldots$ are still an IID sample from Π.



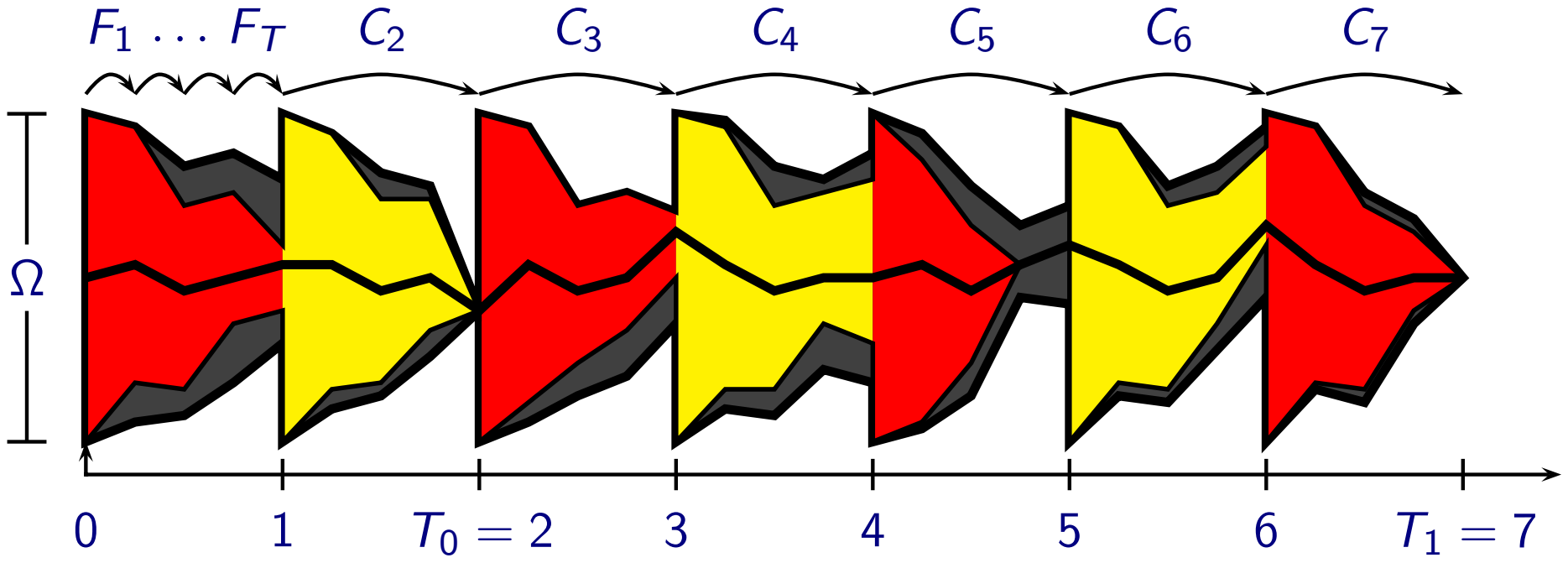
CFTP in Practice

- C is a compound update function $C = F_T \circ \cdots \circ F_1$
- Assume we can determine bounding set $W_t \supseteq C_t(\Omega)$.
- Redefine T_i, so W_{T_i} is "coalescent".
- Then $x_1, x_2, \ldots$ are still an IID sample from Π.



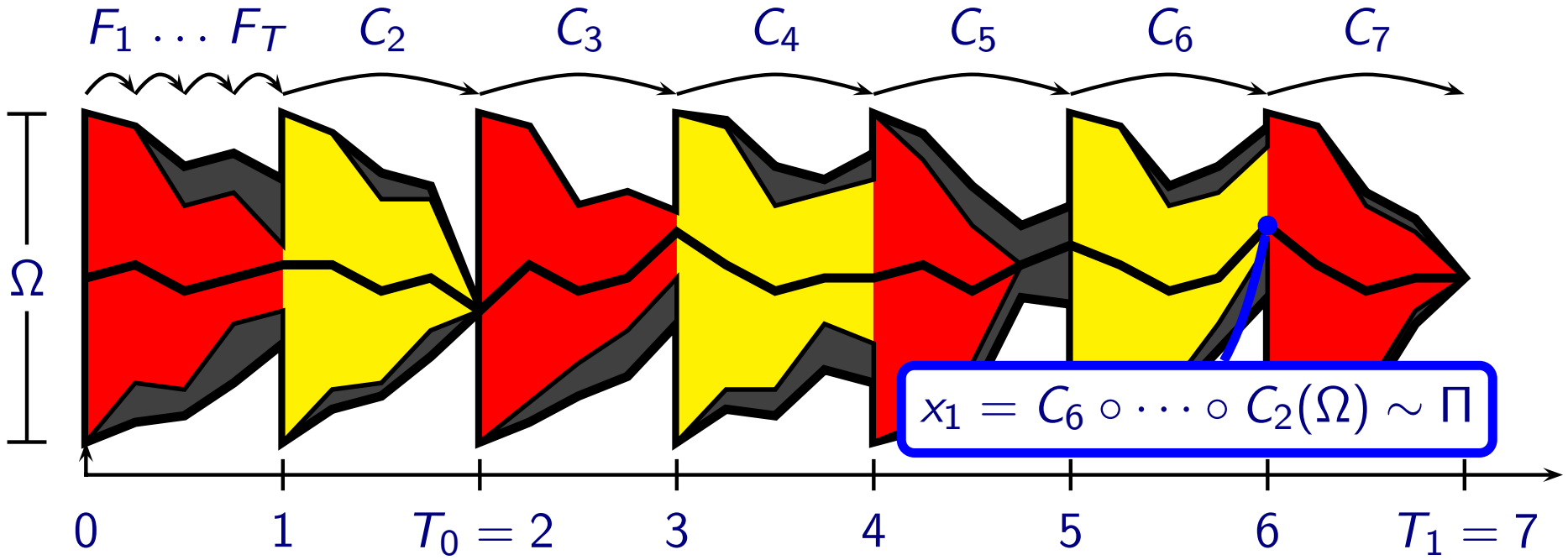
CFTP in Practice

- C is a compound update function $C = F_T \circ \cdots \circ F_1$
- Assume we can determine bounding set $W_t \supseteq C_t(\Omega)$.
- Redefine T_i, so W_{T_i} is "coalescent".
- Then $x_1, x_2, \ldots$ are still an IID sample from Π.



CFTP in Practice

- C is a compound update function $C = F_T \circ \cdots \circ F_1$
- Assume we can determine bounding set $W_t \supseteq C_t(\Omega)$.
- Redefine T_i, so W_{T_i} is "coalescent".
- Then $x_1, x_2, \ldots$ are still an IID sample from Π.



Brute Force Read Once CFTP

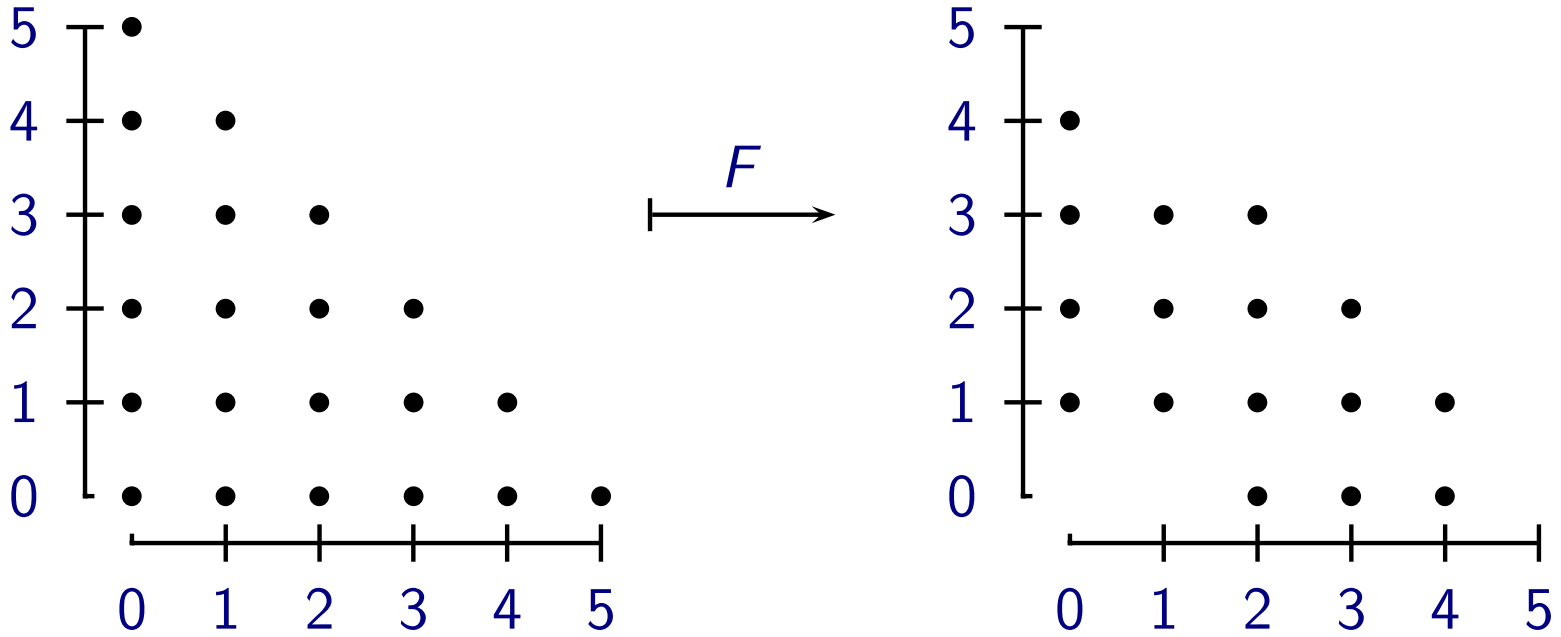
- $C = F_T \circ \cdots \circ F_1$
- Assume that F_1 maps the entire statespace Ω into a finite number of point with probability one.
- Then detecting coalescence is just a question of "tracking" each point in $F_1(\Omega)$ under subsequent mappings $F_2, F_3, \ldots, F_T$.
- Can we do that here?

$F(\Omega)$ *is* Finite

- Recall that $F(\mathbf{x})$ only depends on x through $N_1(\mathbf{x}), \ldots, N_r(\mathbf{x})$.
- Effectively F maps $\mathbb{S}_{r,N} \to \mathbb{S}_{r,N}$, where

$$\mathbb{S}_{r,N} = \{N_1, \ldots, N_{r-1} \in \mathbb{N}_0 : N_1 + \cdots + N_{r-1} \leq N\}.$$

Example: $n = 5$ and $r = 3$



Exact image

- We want to determine $F(\Omega)$.
- Recall:

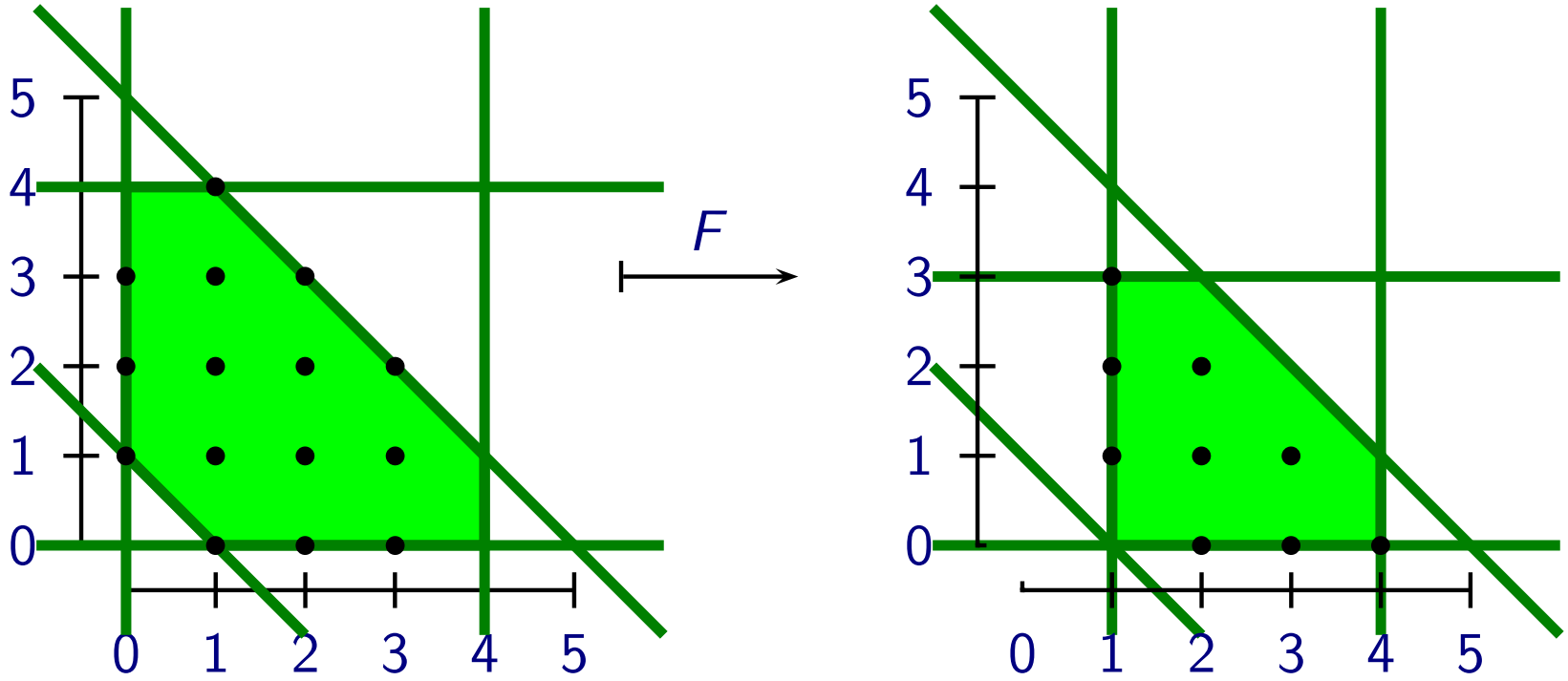
$$M_k(\mathbf{x}) = \frac{G_k(N_k(\mathbf{x}) + 1)}{\sum_{j=1}^{r} G_j(N_j(\mathbf{x}) + 1)}.$$

- Let $L_k = \#\{G_k(i) : i = 1, \ldots, n+1\}$ be the number of unique values taken by G_k.
- Then $\#F(\Omega) \leq \prod_{k=1}^{r} L_k$.
- The complexity of constructing $F(\Omega)$ is $O(rn^{r+1})$.
- Can we do better?

Constructing Bounding Sets

Assume that $\mathbf{x} \in W = \{\mathbf{x} : a_k \leq N_k(\mathbf{x}) \leq b_k\}$.

Given F and W we find $(a'_1, \ldots, a'_r)$ and $(b'_1, \ldots, b'_r)$ so that $\mathbf{x} \in W$ implies $F(\mathbf{x}) \in W'$ where $W' = \{\mathbf{x} : a'_k \leq N_k(\mathbf{x}) \leq b'_k\}$.

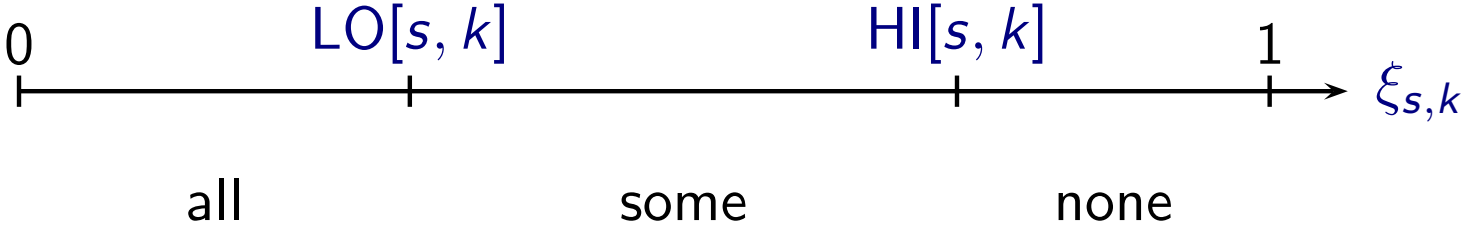


Determining $\mathbf{a}'$ and $\mathbf{b}'$

Assume we have bounds $\mathrm{LO}[s,k]$ and $\mathrm{HI}[s,k]$ so that

$$\mathrm{LO}[s,k] \leq \pi(\eta_s|k)M'_k(x)/\sum_{j=k}^{r}\pi(\eta_s|j)M'_j(x) \leq \mathrm{HI}[s,k]$$

If $Z_s(\mathbf{x}) = j, j < k$ not already accepted for all $\mathbf{x} \in W$, then, for the remaining x, $Z_s(\mathbf{x}) = k$ will be accepted for...



For each s it is now easy to determine a dominating set $D_s \subseteq \{1,\ldots,r\}$, so that $Z_s(\mathbf{x}) \in D_s$ whenever $\mathbf{x} \in W$.

Determining $\mathbf{a}'$ and $\mathbf{b}'$

Given $D_s \subseteq \{1, \ldots, r\}$ we obtain a' and b' by setting

- $a'_k = \#\{s = 1, \ldots, n : D_s = \{k\}\}$
- $b'_k = \#\{s = 1, \ldots, n : \{k\} \subseteq D_s\}$

Constructing each HI and LO has complexity $O(n)$.

Complexity of finding a' and b' is $O(rn^2)$.

Recall: Brute force $= O(rn^{r+1})$.

Calculating $\mathrm{LO}[s,k]$

To find $\mathrm{LO}[s,k]$ we need to find a lower bound on

$$\frac{\pi_k(\eta_s)M'_k(x)}{\sum_{j=k}^{r}\pi_j(\eta_s)M'_j(x)} = \frac{\pi_k(\eta_s)G_k(x)}{\pi_k(\eta_s)G_k(x)+\sum_{j=k+1}^{r}\pi_j(\eta_s)G_j(x)}$$
$$\geq \frac{\pi_k(\eta_s)G_k(a_k+1)}{\pi_k(\eta_s)G_k(a_k+1)+\max_{\mathbf{l}\in\overline{S}(\mathbf{a},\mathbf{b})}\sum_{j=k+1}^{r}\pi_j(\eta_s)G_j(l_j+1)}$$

which follows from the fact that G_i is non-decreasing. The maximum is over those vectors $\mathbf{l}=(l_1,\ldots,l_r)$ belonging to the set

$$\overline{S}(\mathbf{a},\mathbf{b}) := \left\{ l : a_j \leq l_j \leq b_j \text{ for all } j>k \text{ and } \sum_{j=k+1}^{r} l_j \leq n - \sum_{i\leq k} a_i \right\},$$

which is a convex set.

Maximising over $\overline{S}(\mathbf{a}, \mathbf{b})$

Only general way of maximising over $\overline{S}(\mathbf{a}, \mathbf{b})$ is by an exhaustive search which is not feasible.

Instead we construct convex functions $\overline{G}_i$ which bound G_i. Then we have

$$\sum_{j=k+1}^{r} \pi_j(\eta_s) G_j(l_j + 1) \leq \sum_{j=k+1}^{r} \pi_j(\eta_s) \overline{G}_j(l_j + 1). \tag{1}$$

Due to the convexity of $\overline{G}_i$ and $\overline{S}(\mathbf{a}, \mathbf{b})$ it is possible to maximise the RHS of (1) over $\overline{S}(\mathbf{a}, \mathbf{b})$ using a greedy hill climbing algorithm. This maximisation has complexity $O(n)$.

LO$[s,k]$ and HI$[s,k]$

In summary

$$\mathrm{LO}[s,k] = \frac{\pi_k(\eta_s)G_k(a_k+1)}{\pi_k(\eta_s)G_k(a_k+1) + \max_{l\in\overline{S}(\mathbf{a},\mathbf{b})}\sum_{j=k+1}^{r}\pi_j(\eta_s)\overline{G}_j(l_j+1)}.$$

Using similar arguments we obtain

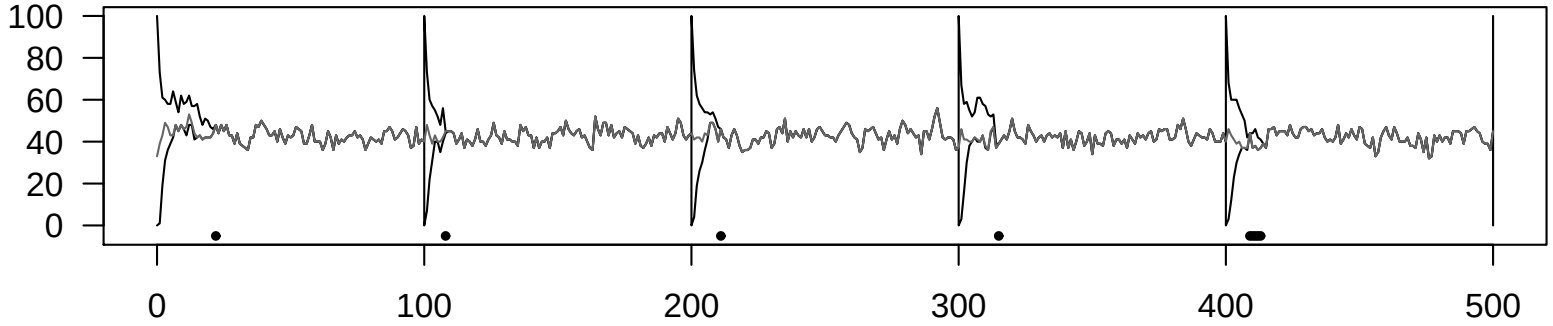
$$\mathrm{HI}[s,k] = \frac{\pi_k(\eta_s)G_k(b_k+1)}{\pi_k(\eta_s)G_k(b_k+1) + \min_{\mathbf{l}\in\underline{S}(\mathbf{a},\mathbf{b})}\sum_{j=k+1}^{r}\pi_j(\eta_s)\underline{G}_j(l_j+1)},$$

where $\underline{S}(\mathbf{a},\mathbf{b})$ is similar to $\overline{S}(\mathbf{a},\mathbf{b})$ and each function $\underline{G}_i$ is a concave lower bound on G_i which allow a greedy minimisation.

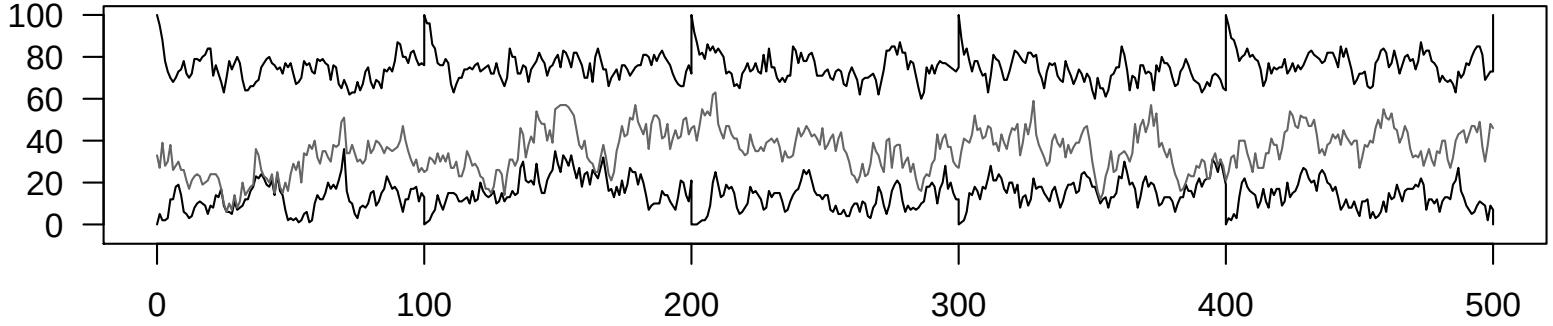
Example

We assume normal mixture components $\pi(\cdot|k) = \mathcal{N}(\mu_k, \sigma^2)$.

Plot shows trace of $a_1$and b_1 when $n = 100$, $r = 3$, $\mu = (0, 1, 2)$, $m = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $\sigma^2 = 0.25$.



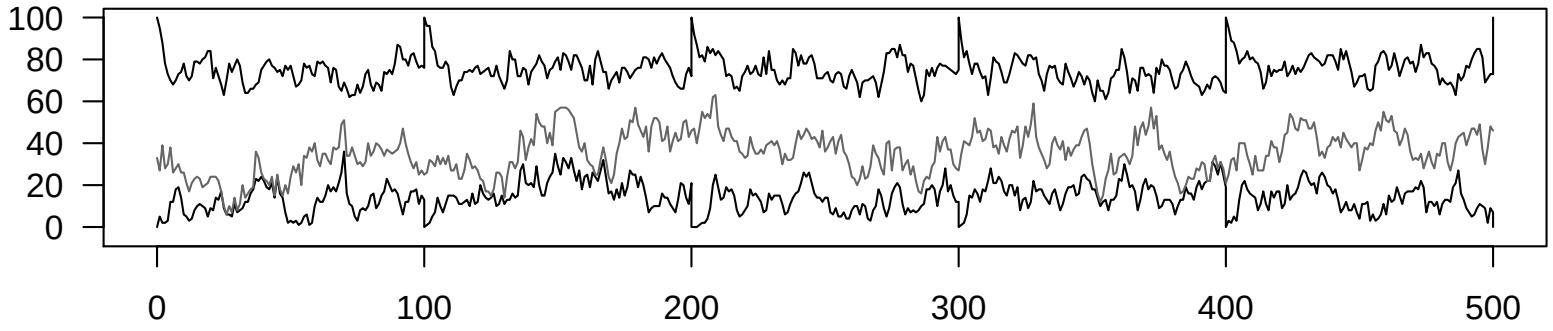
As above but with $\mu = (0, 0.5, 1)$ — no coalescence.



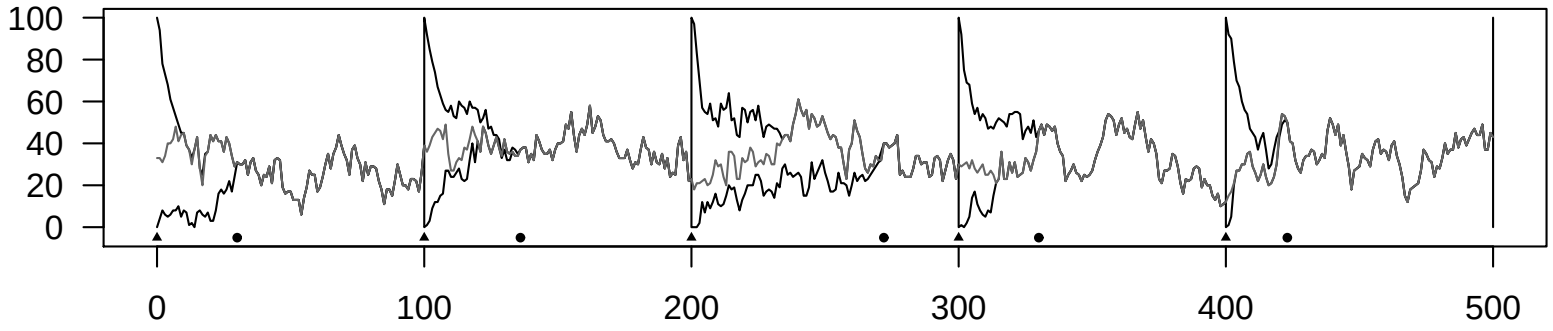
Example - cont.

Plot shows trace of a_1 and b_1 when $n = 100$, $r = 3$, $\mu = (0, 0.5, 1)$, $m = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $\sigma^2 = 0.25$.

Using bounding sets. (Plot from before)



Using brute force.



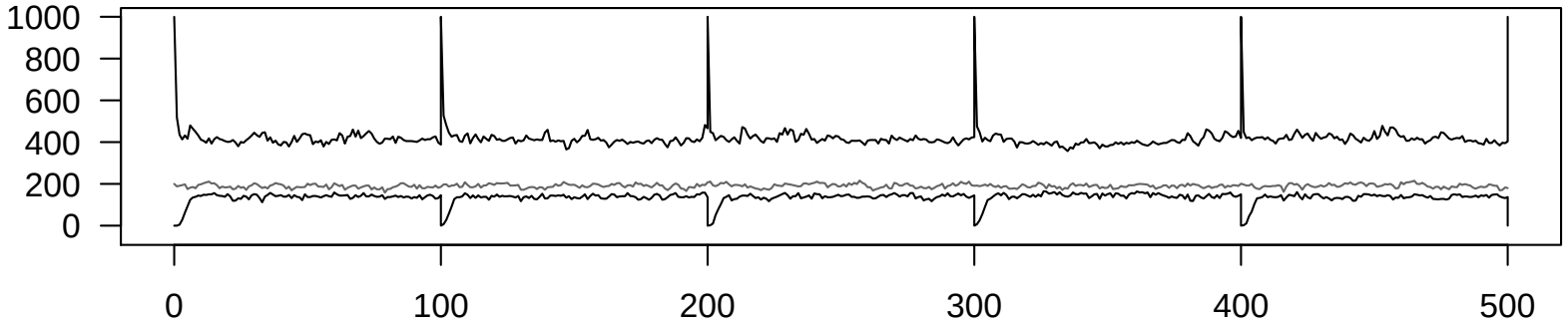
Summary

- Bounding sets:
  - For: Low complexity, $O(rn^2)$.
  - Against: Sloppy bounding sets.
- Brute force CFTP:
  - For: "Exact" bounding sets.
  - Against: High complexity, $O(rn^{r+1})$.

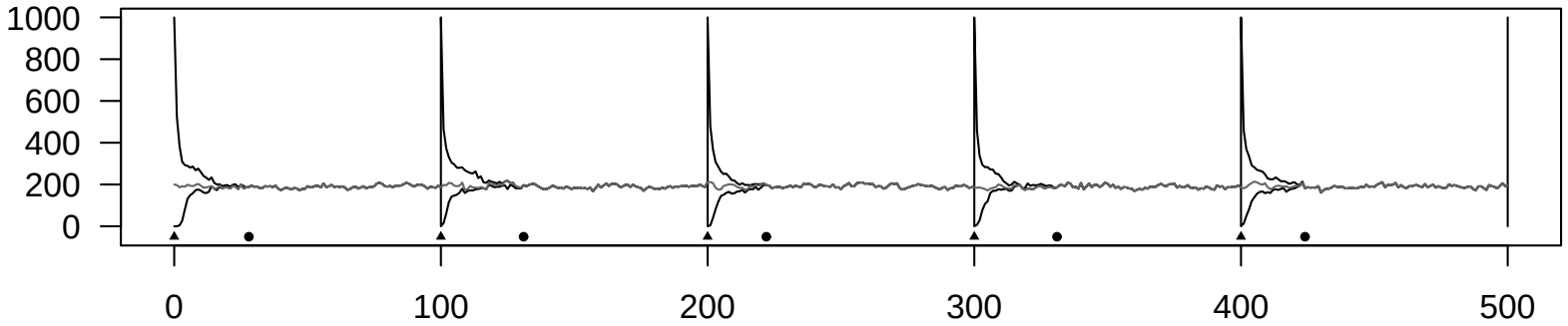
- **Idea**: Use bounding sets until "volume" of the bounding set $\prod_{i=1}^{r}(b_i - a_1 + 1)$ is below some threshold. When the volume is small enough we deem brute force to be "cheap enough".

- **Example**: Next slide, $n = 1000$, $r = 5$, $\mu = (0, 1, 2, 3, 4)$, $m = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $\sigma^2 = 0.25$.

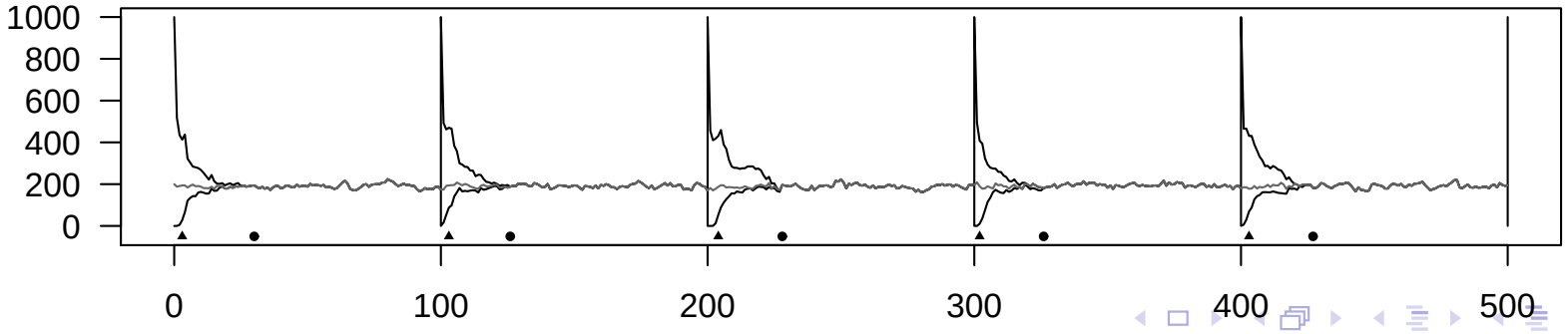
Bounding sets:



Brute force:



Bounding sets until "volume" is below exp(30).



Results

Average time per sample.

- Bounding sets:
  - NA (never coalesces).
- Brute force CFTP:
  - 72.79 sec/sample.
- Bounding sets until "volume" is below exp(30):
  - 22.75 sec/sample.

Conclusion:
Perfect sampling of the posterior weights is available.

Furthermore, the basic algorithm can be greatly improved by applying brute force at "the right time".

- Extend to case of unknown component means
- General problem: label switching
- Here $F(\mathbf{x})$ depends on $\mathbf{x}$ through the exact configuration of $\mathbf{z}$ and not only $N(x)$.
- Approach is feasible for $n = 5$ data points...