Short Course

State Space Models, Generalized Dynamic Systems
and
Sequential Monte Carlo Methods,
and
their applications
in Engineering, Bioinformatics and Finance

Rong Chen
Rutgers University
Peking University

Part Three: Advanced Sequential Monte Carlo

3.1 Mixture Kalman Filter

3.1.1 Conditional Dynamic Linear Models

3.1.2 Mixture Kalman Filters

3.1.3 Partial Conditional Dynamic Linear Models

3.1.4 Extend Mixture Kalman Filters

3.1.5 Future Directions

3.2 Constrained SMC

3.3 Parameter Estimation in SMC

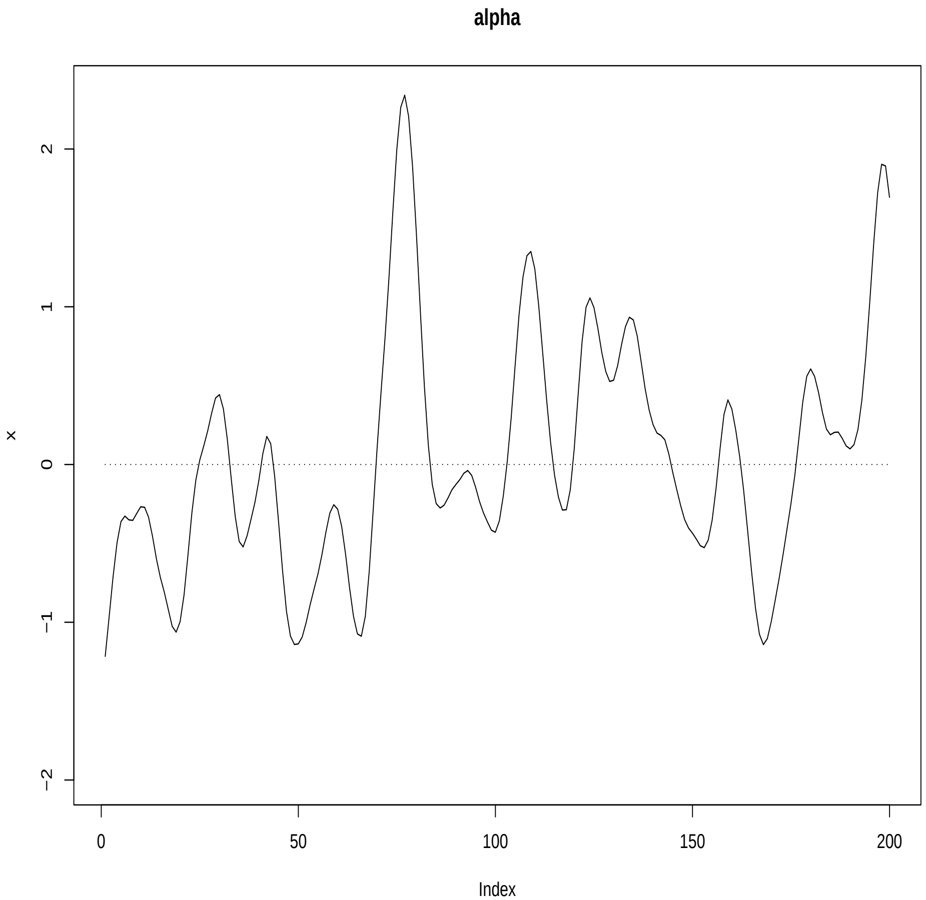
3.4 Look-Ahead Strategies

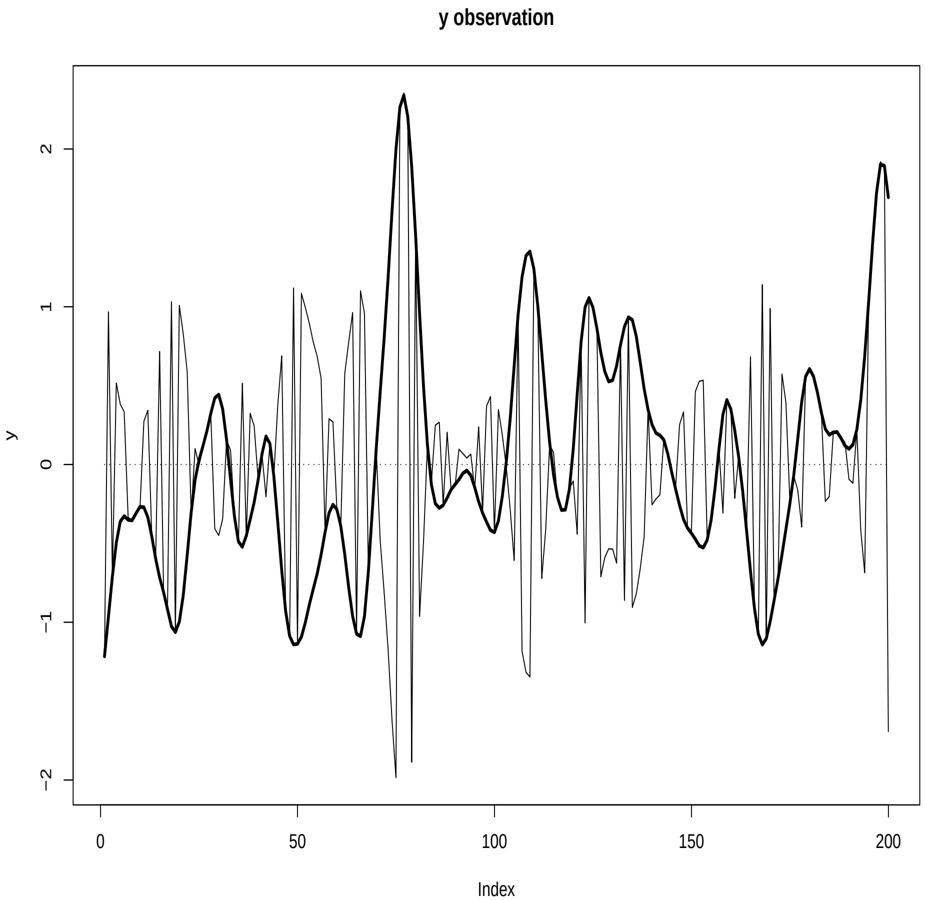
3.4.1. The principle of lookahead

- **Dynamic systems often process strong 'memory'**
- **Future observations can reveal substantial information on the current state**
- **Slight delay is tolerable**

Make inference on the state x_t at time $t+d$, based on observations

$$y_1, \ldots, y_t, y_{t+1}, \ldots, y_{t+d}.$$


alpha
x
2
1
0
−1
−2
0
50
100
150
200
Index


y observation
y
2
1
0
−1
−2
0
50
100
150
200
Index

If $\hat{h}_{t+d}$ is a consistent MC estimator of $E(h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d})$, then

$$
\begin{aligned}
E\left[\hat{h}_{t+d} - h(\boldsymbol{x}_t)\right]^2 \;=\; & E\left[\hat{h}_{t+d} - E(h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d})\right]^2 \\
& + E\left[E(h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d}) - h(\boldsymbol{x}_t)\right]^2
\end{aligned}
$$

- **The first term goes to zero with as MC sample size increases**
- [*Proposition*] **The second term decreases as d increases**
- **When MC sample size is sufficiently large, the first term is negligible comparing to the second term, then longer lookahead (larger d) always improves efficiency**
- **With limited sample size and limited computational time, lookahead may not always be more efficient.**

3.4.2. Lookahead algorithms

(1) Exact lookahead weighting

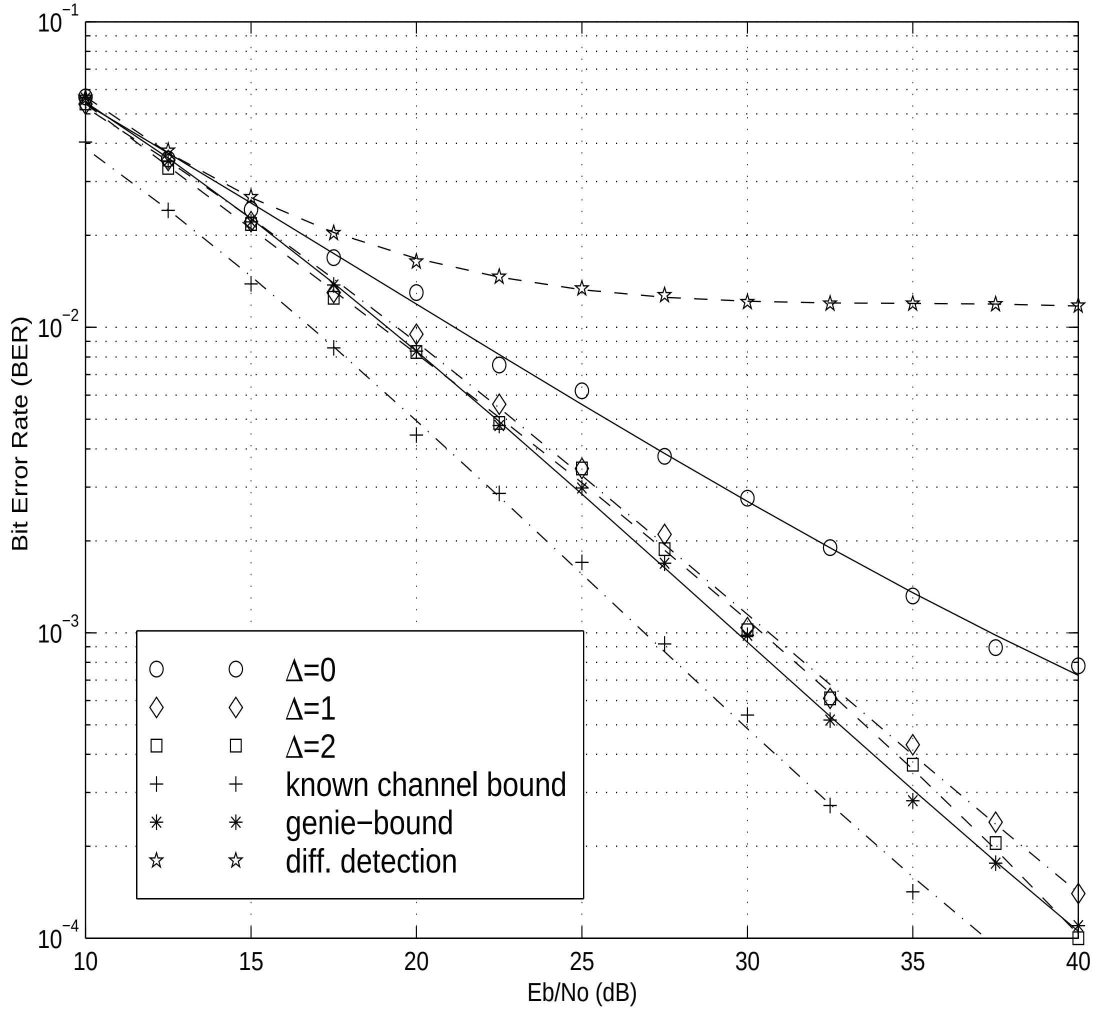
If: $(\boldsymbol{x}_{t+d}^{(j)}, w_{t+d}^{(j)})$ **is properly weighted w.r.t.** $p(\boldsymbol{x}_{t+d} \mid \boldsymbol{y}_{t+d})$

then: $(x_t^{(j)}, w_{t+d}^{(j)})$ **is properly weighted w.r.t.** $p(x_t \mid \boldsymbol{y}_{t+d})$**.**

Hence, inference on x_t **can be made using** $(x_t^{(j)}, w_{t+d}^{(j)})$**, with concurrent SMC.**

- **Sample of** x_t **is drawn at time** t**, based on** $\boldsymbol{y}_t$**, with weight** w_t
- **Inference on** x_t **is made at time** $t+d$**, with weight** w_{t+d}**;**
  —— w_{t+d} **is based on** $\boldsymbol{y}_{t+d}$ **and samples of** $(\boldsymbol{x}_t, x_{t+1}, \ldots, x_{t+d})$**.**

$$E(h(x_t) \mid \boldsymbol{y}_{t+d}) \approx \frac{\sum_{j=1}^m h(x_t^{(j)}) w_{t+d}^{(j)}}{\sum_{j=1}^m w_{t+d}^{(j)}}$$



10^{-1}
10^{-2}
10^{-3}
10^{-4}
Bit Error Rate (BER)
10
15
20
25
30
35
40
Eb/No (dB)
Δ=0
Δ=1
Δ=2
known channel bound
genie−bound
diff. detection

In fact:

$$w_{t+d}^{(j)} = w_{t-1}^{(j)} \frac{\pi_{t+d}(\boldsymbol{x}_{t+d}^{(j)})}{\pi_{t-1}(\boldsymbol{x}_{t-1}^{(j)}) \prod_{s=t}^{t+d} g_s(x_s^{(j)} \mid \boldsymbol{x}_{s-1}^{(j)}, \boldsymbol{y}_s)}$$

An improved version (if practical) is to use

$$\tilde{w}_{t+d}^{(j)} = w_{t-1}^{(j)} \frac{\pi_{t+d}(\boldsymbol{x}_t^{(j)})}{\pi_{t-1}(\boldsymbol{x}_{t-1}^{(j)}) g_t(x_t^{(j)} \mid \boldsymbol{x}_{t-1}^{(j)}, \boldsymbol{y}_t)}$$

where $\pi_{t+d}(\boldsymbol{x}_t^{(j)}) = \int \pi_{t+d}(\boldsymbol{x}_t^{(j)}, x_{t+1}, \ldots, x_{t+d}) dx_{t+1} \ldots dx_{t+d}$

$[Proposition]$**:**

$$Var\left[w_{t+d} \mid \boldsymbol{y}_{t+d}\right] \geq Var\left[\tilde{w}_{t+d} \mid \boldsymbol{y}_{t+d}\right]$$

and

$$Var\left[w_{t+d} h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d}\right] \geq Var\left[\tilde{w}_{t+d} h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d}\right]$$

(2) Exact Lookahead Sampling

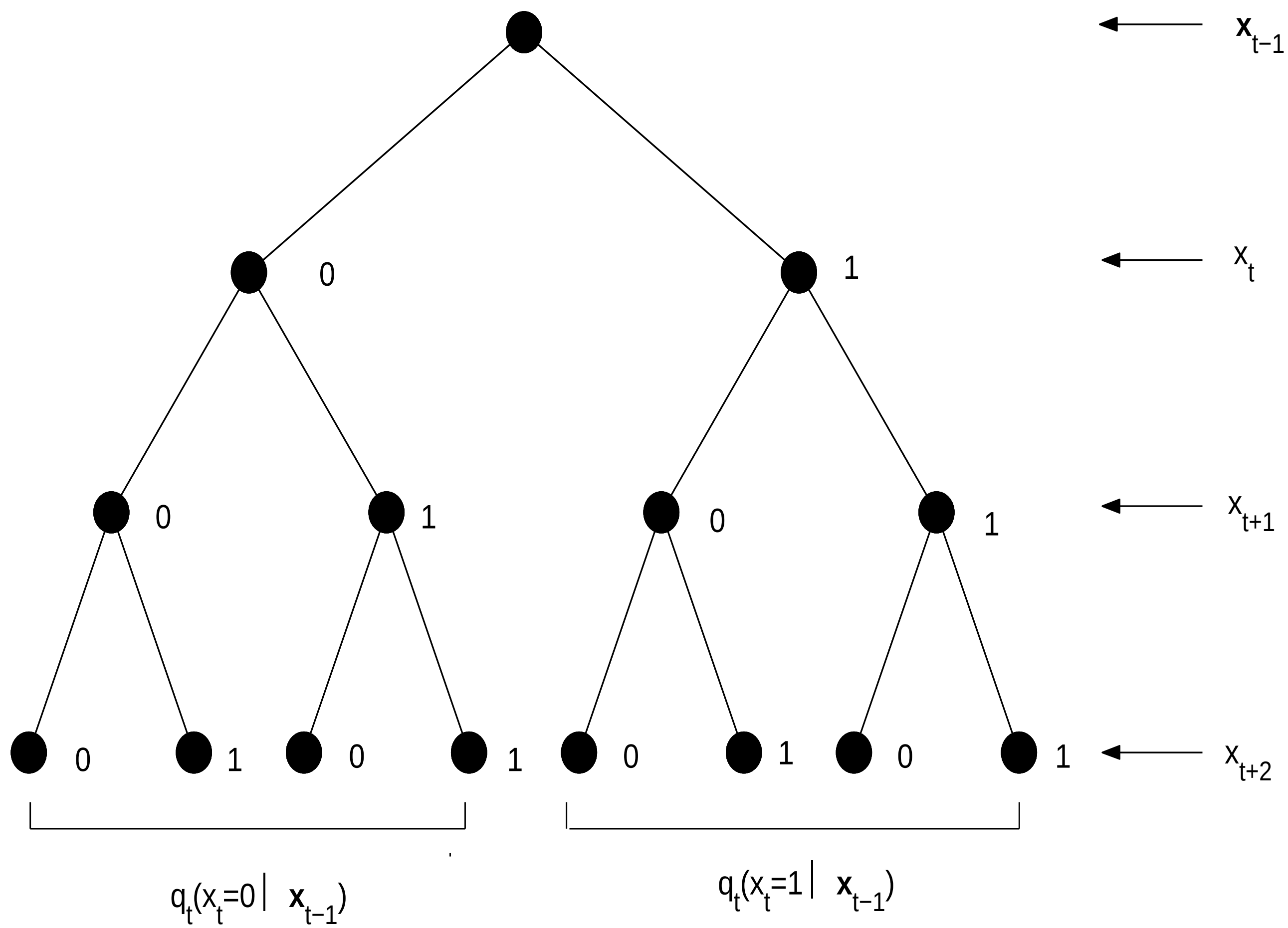
- **New target distribution:** $\pi_t^*(\boldsymbol{x}_t) = p(\boldsymbol{x}_t \mid \boldsymbol{y}_{t+d})$
- **— recall: concurrent target distribution:** $\pi_t(\boldsymbol{x}_t) = p(\boldsymbol{x}_t \mid \boldsymbol{y}_t)$
- **Sample** x_t **based on a trial distribution that uses the full information** $\boldsymbol{y}_{t+d}$

In particular, we can use

$$
\begin{aligned}
g_t(x_t \mid \boldsymbol{x}_{t-1}^{(j)}) &= p(x_t \mid \boldsymbol{x}_{t-1}^{(j)}, \boldsymbol{y}_{t+d}) \\
&= \int p(x_t, x_{t+1}, \ldots, x_{t+d} \mid \boldsymbol{x}_{t-1}^{(j)}, \boldsymbol{y}_{t+d}) dx_{t+1} \ldots x_{t+d}
\end{aligned}
$$

Then

$$
w_t^{*(j)} \propto w_{t-1}^{*(j)} \frac{\pi_t^*(\boldsymbol{x}_t)}{\pi_{t-1}^*(\boldsymbol{x}_{t-1}) g_t(x_t \mid \boldsymbol{x}_{t-1})} = w_{t-1}^{*(j)} \frac{p(\boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{t+d})}{p(\boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{t+d-1})}
$$



$\mathbf{x}_{t-1}$
x_t
x_{t+1}
x_{t+2}
0
1
0
1
0
1
0
1
0
1
0
1
0
1
$q_t(x_t=0 \mid \mathbf{x}_{t-1})$
$q_t(x_t=1 \mid \mathbf{x}_{t-1})$

We compare the (improved) exact lookahead weighting with the exact lookahead sampling methods:

Suppose at time t**,** $(\boldsymbol{x}_t^{(j)}, w_t^{(j)})$ **properly weighted w.r.t.** $\pi_t(\boldsymbol{x}_t) = p(\boldsymbol{x}_t \mid \boldsymbol{y}_t)$**.**

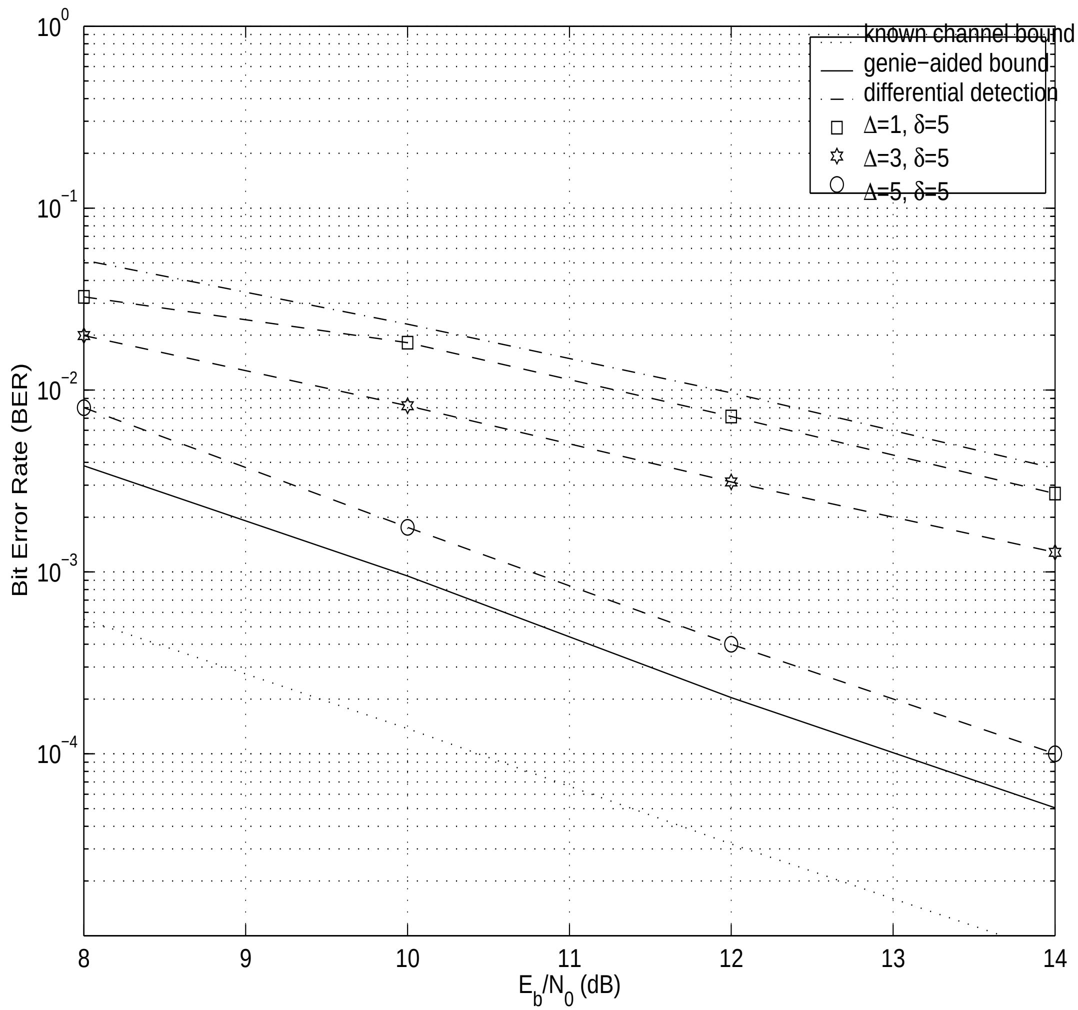
$[Proposition]$

$$Var[\tilde{w}_{t+d} \mid \boldsymbol{y}_{t+d}] \geq Var[w_t^* \mid \boldsymbol{y}_{t+d}]$$

and

$$Var[\tilde{w}_{t+d} h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d}] \geq Var[w_t^* h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d}]$$

However: Excessive computing cost

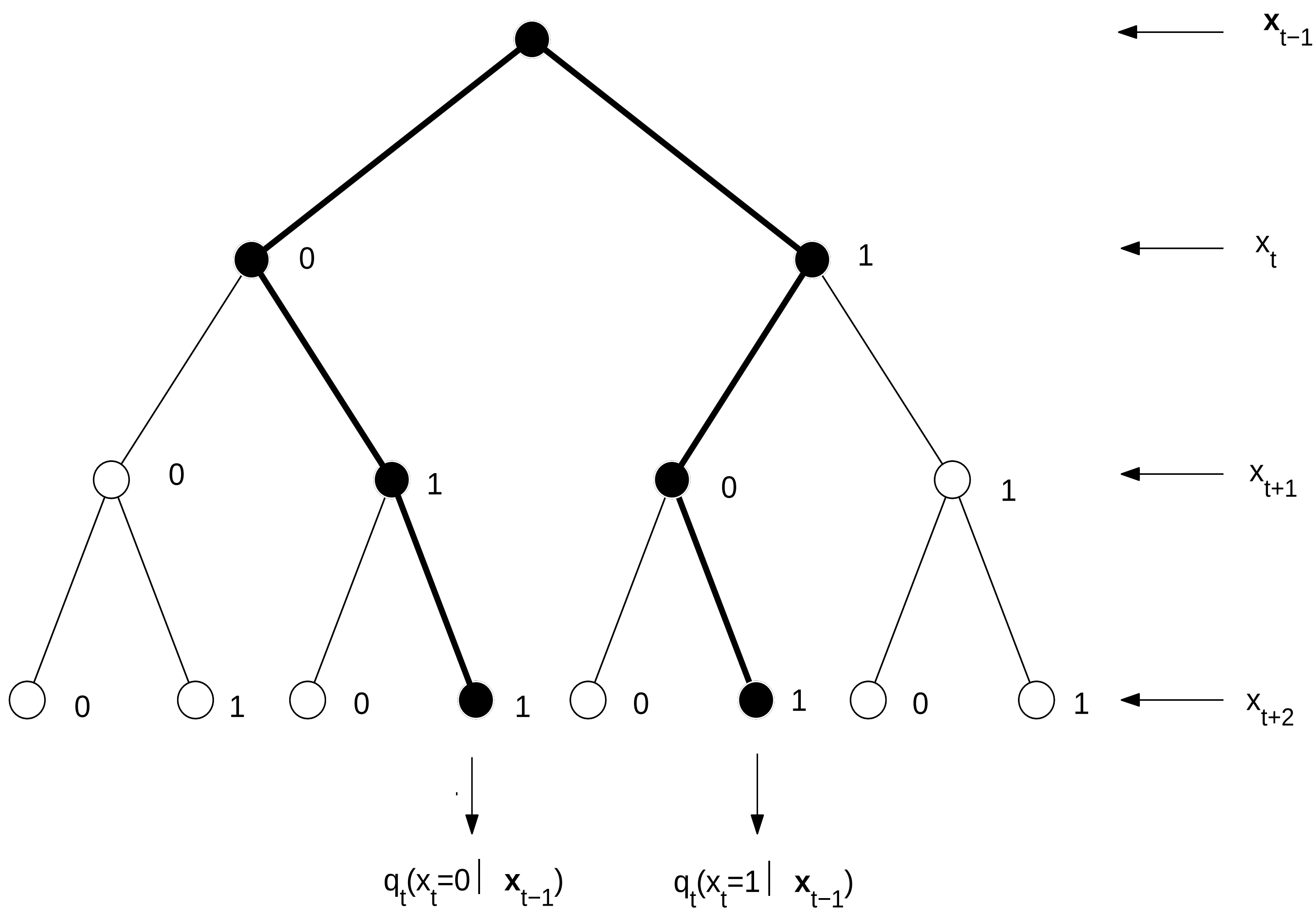

known channel bound
genie−aided bound
differential detection
Δ=1, δ=5
Δ=3, δ=5
Δ=5, δ=5
Bit Error Rate (BER)
E_b/N_0 (dB)
10^0
10^−1
10^−2
10^−3
10^−4
8
9
10
11
12
13
14

(3) Pilot lookahead sampling:

- **Full exploration of future space is too expensive**
- **Pilots can be used to partially explore the future space**
- **Limited numbers of pilots are able to obtain useful future information with low computational cost**

Specifically,

- **Suppose** x_t **takes** J **possible values** $\{a_1, \ldots, a_J\}$
- **Starting with each possible** $x_t = a_i$**, propagate to** x_{t+d} **with concurrent SMC with optimal sampling distribution.**
- **Obtain pilot incremental weight for each pilot**
- **Sample** x_t **from** $\{a_1, \ldots, a_J\}$ **according to pilot incremental weight**
- **Update weight**



0
1
0
1
0
1
0
1
0
1
0
1
0
1
$\mathbf{x}_{t-1}$
x_t
x_{t+1}
x_{t+2}
$q_t(x_t=0 \mid \mathbf{x}_{t-1})$
$q_t(x_t=1 \mid \mathbf{x}_{t-1})$

More specifically

- **For each $\boldsymbol{x}_{t-1}^{(j)}$ and each a_i,**
- **generate $x_{t+1}, \ldots, x_{t+d}$ from**

$$\prod_{s=t+1}^{t+d} \pi_s(x_s \mid \boldsymbol{x}_{t-1}^{(j)}, a_i, x_{t+1}, \ldots, x_{s-1})$$

- **Obtain the pilot incremental weight**

$$U_t^{(i,j)} = \frac{\pi_{t+d}(\boldsymbol{x}_{t-1}^{(j)}, x_t = a_i, x_{t+1}^{(i,j)}, \ldots, x_{t+d}^{(i,j)})}{\pi_{t-1}(\boldsymbol{x}_{t-1}^{(j)}) \prod_{s=t+1}^{t+d} \pi_s(x_s^{(i,j)} \mid \boldsymbol{x}_{t-1}^{(j)}, x_t = a_i, x_{t+1}^{(i,j)}, \ldots, x_{s-1}^{(i,j)})}$$

- **sample $x_t^{(j)}$ from $\{a_1, \ldots, a_J\}$ with probability**

$$g_t(x_t = a_i \mid \boldsymbol{x}_{t-1}^{(j)}) = \frac{U_t^{(i,j)}}{\sum_{k=1}^{J} U_t^{(k,j)}}$$

- **New weight**

$$w_t^{(j)} = w_{t-1}^{(j)} \frac{\pi_t(\boldsymbol{x}_t^{(j)})}{\pi_{t-1}(\boldsymbol{x}_{t-1}^{(j)}) g_t(x_t = x_t^{(j)} \mid \boldsymbol{x}_{t-1})}$$

Note:

- (x_t, w_t) **is properly weighted w.r.t** $p(x_t \mid \boldsymbol{y}_t)$**.**
- **even though sampling is done with future information**
- **Inference using** (x_t, w_t) **is not efficient**

$$\frac{\sum h(x_t^{(j)})w_t^{(j)}}{\sum w_t^{(j)}} \approx E(h(x_t) \mid \boldsymbol{y}_t)$$

Remedies:

- **To make inference on** x_t **at time** $t+d$**, calculate**

$$w_{t+d}^{**(j)} = w_{t-1}^{(j)} \sum_{i=1}^{J} U_t^{(i,j)}$$

- **Then** $(x_t^{(j)}, w_{t+d}^{**(j)})$ **is properly weighted w.r.t.** $p(x_t \mid \boldsymbol{y}_{t+d})$**.**
- **To make inference on** x_t **at time** $t+d$**, use** $(x_t^{(j)}, w_{t+d}^{**(j)})$

Further improvement: Multi-pilot lookahead sampling: using multiple pilots per a_i.

Comparison with exact lookahead sampling:

[*Proposition*]

$$0 \le Var(w_t^{**}) - Var(w_t^*) \sim O(1/K)$$

where K is number of pilots per a_i and w_t^* is the weight of exact lookahead sampling.

Recall:

$$Var(w_t^*) \ge Var(\tilde{w}_t)$$

where $\tilde{w}_t$ is the weight of improved delay weighting algorithm.

Deterministic pilot lookahead

- **Pilots need not be random**
- **A better pilot might be the path that maximize**

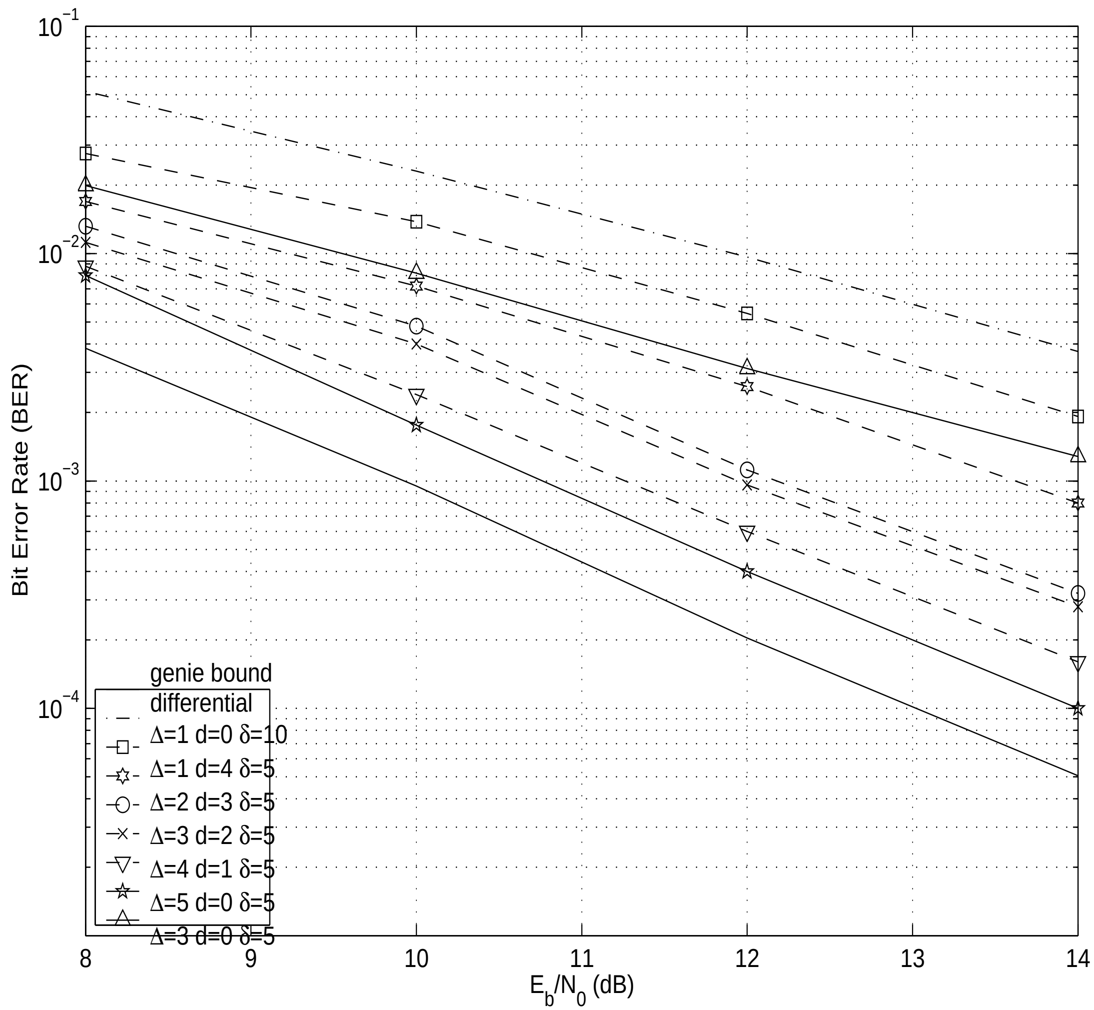
$$\pi_{t+d}(x_{t+1}, \ldots, x_{t+d} \mid \boldsymbol{x}_{t-1}^{(j)}, x_t = a_i)$$

- **The true maximum is too expensive to get**
- **A greedy sequential search: for** $s = t+1, \ldots, t+d$

$$x_s^{(i,j)} = \arg\max_{x_s} \pi_s(x_s \mid \boldsymbol{x}_{t-1}^{(j)}, x_t = a_i, x_{t+1}^{(i,j)}, \ldots, x_{s-1}^{(i,j)})$$

- $$g_t(x_t = a_i \mid \boldsymbol{x}_{t-1}^{(j)}) \propto \frac{\pi_{t+d}(\boldsymbol{x}_{t-1}^{(j)}, x_t = a_i, x_{t+1}^{(i,j)}, \ldots, x_{t+d}^{(i,j)})}{\pi_{t-1}(\boldsymbol{x}_{t-1}^{(j)})}$$

- **Often, the samples are better than (random) one-pilot lookahead sampling.**



Bit Error Rate (BER)
E_b/N_0 (dB)
genie bound
differential
Δ=1 d=0 δ=10
Δ=1 d=4 δ=5
Δ=2 d=3 δ=5
Δ=3 d=2 δ=5
Δ=4 d=1 δ=5
Δ=5 d=0 δ=5
Δ=3 d=0 δ=5

(4) Adaptive Sampling

- **Sample from a simple trial distribution when information is strong.**
- **Sample from a better trial distribution (e.g. lookahead) when information is weak.**

Recall:

$$E\left[\hat{h}_{t+d} - h(\boldsymbol{x}_t)\right]^2 = E\left[\hat{h}_{t+d} - E(h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d})\right]^2 + E\left[E(h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d}) - h(\boldsymbol{x}_t)\right]^2$$

With finite MC samples, the first term may increase as d increases.

Exact lookahead weighting: comparing $t+d-1$ and $t+d$

$$\hat{h} = \frac{1}{m}\sum_{j=1}^{m} w_{t+d-1}^{(j)} h(\boldsymbol{x}_t^{(j)}) \rightarrow E(h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d-1})$$

$$\tilde{h} = \frac{1}{m}\sum_{j=1}^{m} w_{t+d}^{(j)} h(\boldsymbol{x}_t^{(j)}) \rightarrow E(h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d})$$

$[Proposition]$**: when**

$$E\left[Var\left(\tilde{h} \mid \boldsymbol{x}_{t+d-1}^{(1:m)}, \boldsymbol{y}_{t+d-1}\right) \mid \boldsymbol{y}_{t+d-1}\right] \geq 2Var\left[E\left(h(\boldsymbol{x}_t) \mid \boldsymbol{y}_{t+d}\right) \mid \boldsymbol{y}_{t+d-1}\right]$$

we have

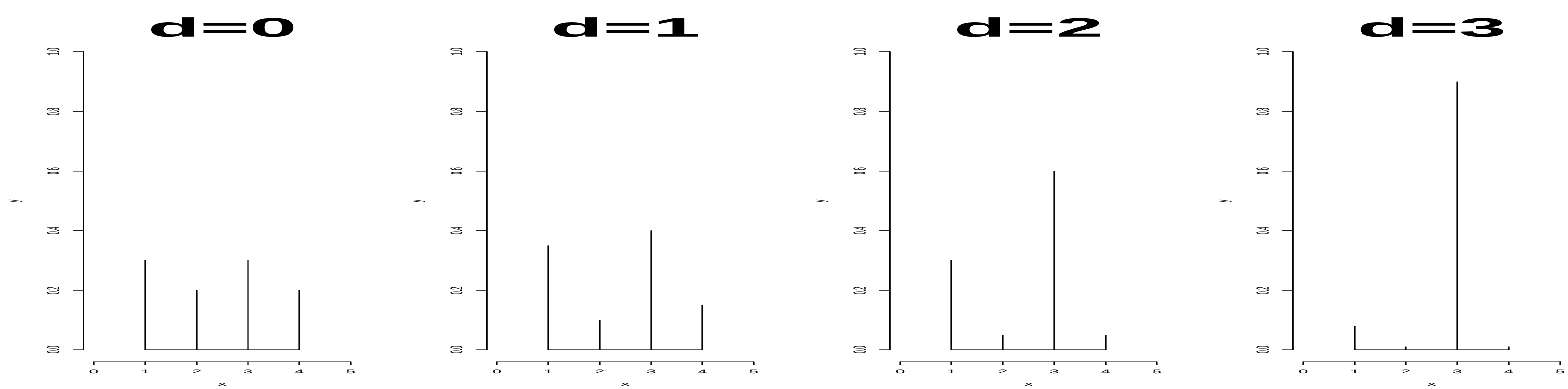
$$E\left[\left(\hat{h} - h(\boldsymbol{x}_t)\right)^2 \mid \boldsymbol{y}_{t+d-1}\right] \geq E\left[\left(\tilde{h} - h(\boldsymbol{x}_t)\right)^2 \mid \boldsymbol{y}_{t+d-1}\right].$$

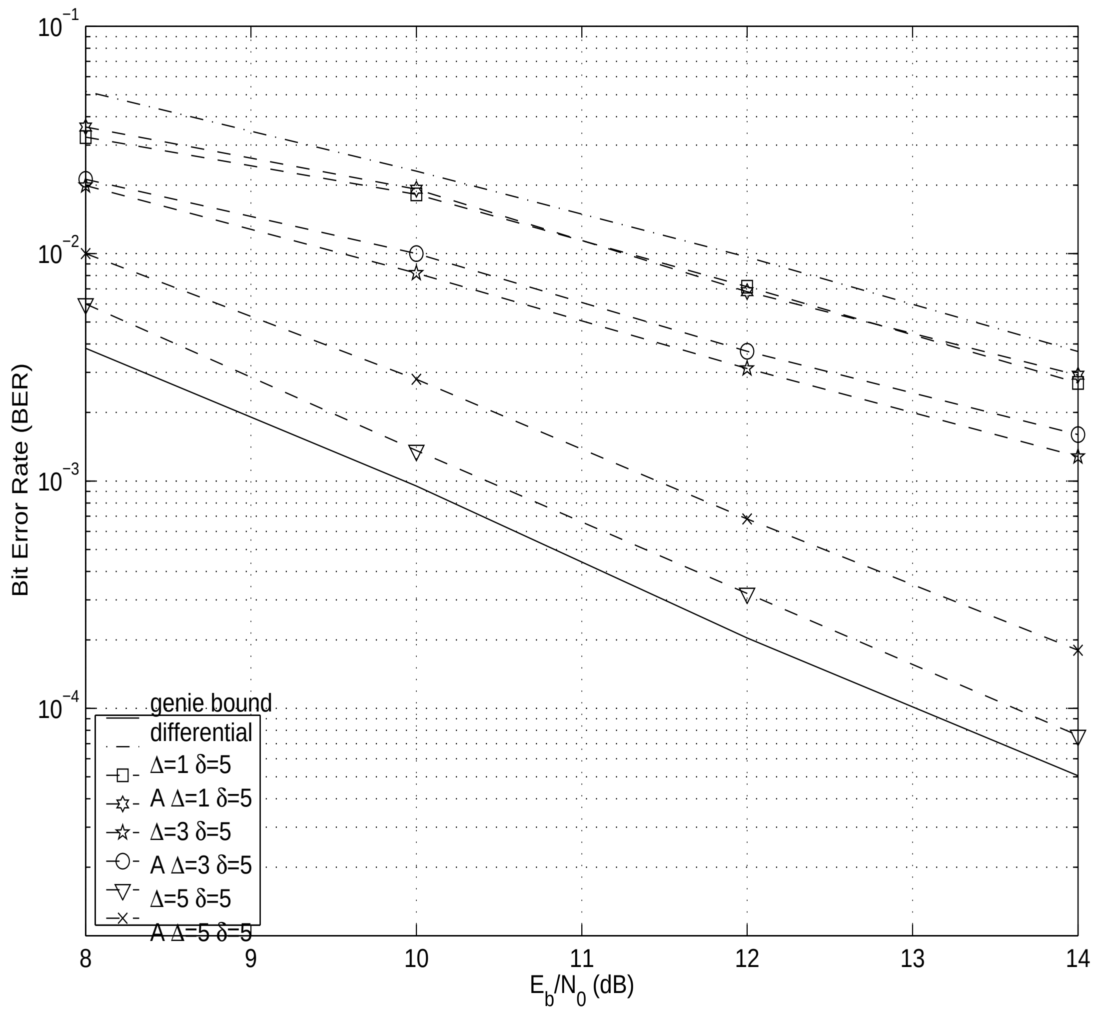
Remarks:

- When $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{t+d-1}) = p(\boldsymbol{x}_t \mid \boldsymbol{y}_{t+d})$, the result holds.
- In general, the condition is difficult to check
- Instead, we check if the information is strong:
  - Specifically, iteratively try $d = 1, 2, \ldots, d_{max}$. Stop when

$$max_{i_0}\Big\{\widehat{\pi}_{t+d}(x_t = a_{i_0})\Big\} \doteq max_{i_0}\Big\{\frac{\sum_j w_{t-1}^{(j)} U_t^{(i_0,j)}}{\sum_{i,j} w_{t-1}^{(j)} U_t^{(i,j)}}\Big\} > p_0,$$

for $p_0 > 0$ but close to 1. (discrete state space)




Bit Error Rate (BER)
E_b/N_0 (dB)
genie bound
differential
Δ=1 δ=5
A Δ=1 δ=5
Δ=3 δ=5
A Δ=3 δ=5
Δ=5 δ=5
A Δ=5 δ=5

Other applications:

- **Multi-target tracking in clutter**
- **Self-avoiding walks modelling protein structure**
- **Generating samples of diffusion bridges**
- **Signal processing in more complex fading channels**