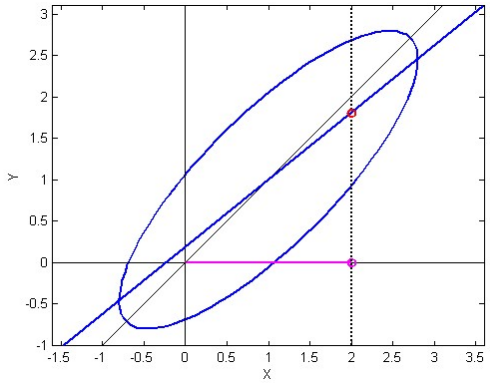
Machine Learning for Signal Processing
Predicting and Estimation from Time Series: Part 2

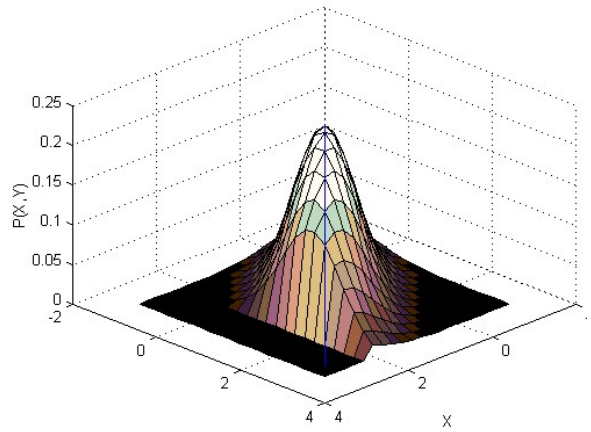
Bhiksha Raj

Preliminaries : P(y|x) for Gaussian

- If P(x,y) is Gaussian:

$$P(\mathbf{x},\mathbf{y}) = N(\begin{bmatrix}\mu_{\mathbf{x}}\\ \mu_{\mathbf{y}}\end{bmatrix}, \begin{bmatrix}C_{\mathbf{xx}} & C_{\mathbf{xy}}\\ C_{\mathbf{yx}} & C_{\mathbf{yy}}\end{bmatrix})$$





- The conditional probability of y given x is also Gaussian
  - The slice in the figure is Gaussian

$$P(y \mid x) = N(\mu_y + C_{yx}C_{xx}^{-1}(x - \mu_x), C_{yy} - C_{yx}C_{xx}^{-1}C_{xy})$$

- The mean of this Gaussian is a function of x
- The variance of y reduces if x is known
  - Uncertainty is reduced

Background: Sum of Gaussian RVs

$$O = AS + \varepsilon$$

$$S \sim N(\mu_S, \Theta_S) \qquad \varepsilon \sim N(\mu_\varepsilon, \Theta_\varepsilon)$$

- The conditional probability of *O*:

$$P(O|S) = N(AS + \mu_\varepsilon, \Theta_\varepsilon)$$

- The overall probability of *O*:

$$P(O) = N(A\mu_S + \mu_\varepsilon, A\Theta_S A^{\mathrm{T}} + \Theta_\varepsilon)$$

Background: Joint Prob. of O and S

$$O = AS + \varepsilon$$

$$Z = \begin{bmatrix} O \\ S \end{bmatrix}$$

- The joint probability of *O* and *S* (i.e. P(*Z*)) is also Gaussian

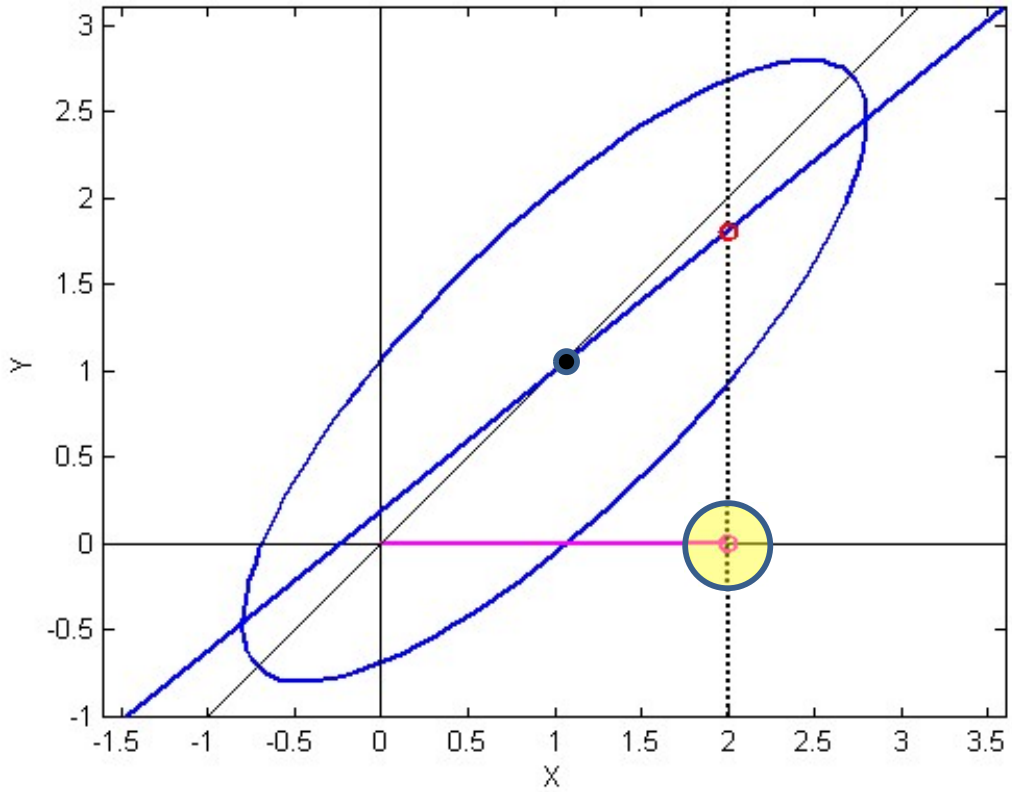
$$P(Z) = P(O, S) = N(\mu_Z, \Theta_Z)$$

- Where

$$\mu_Z = \begin{bmatrix} \mu_O \\ \mu_S \end{bmatrix} = \begin{bmatrix} A\mu_s + \mu_\varepsilon \\ \mu_S \end{bmatrix}$$

- $$\Theta_Z = \begin{bmatrix} \Theta_O & \Theta_{OS} \\ \Theta_{SO} & \Theta_S \end{bmatrix} = \begin{bmatrix} A\Theta_S A^T + \Theta_\varepsilon & A\Theta_S \\ \Theta_S A^T & \Theta_S \end{bmatrix}$$

Preliminaries : Conditional of S given O: P(S|O)

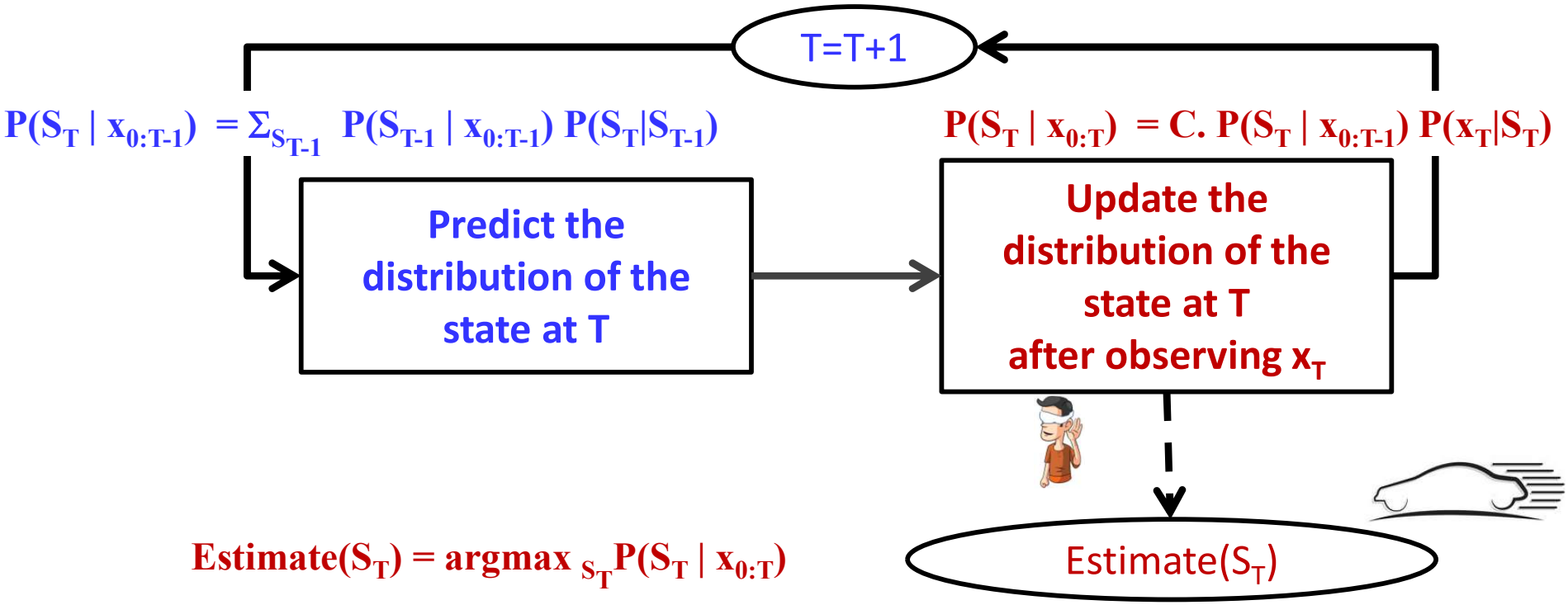


$$\boldsymbol{O = AS + \varepsilon}$$

$$\boldsymbol{P(S|O) = N(\mu_S + \Theta_{SO}\Theta_O^{-1}(O - \mu_O),\ \ \Theta_S - \Theta_{SO}\Theta_O^{-1}\Theta_{OS})}$$

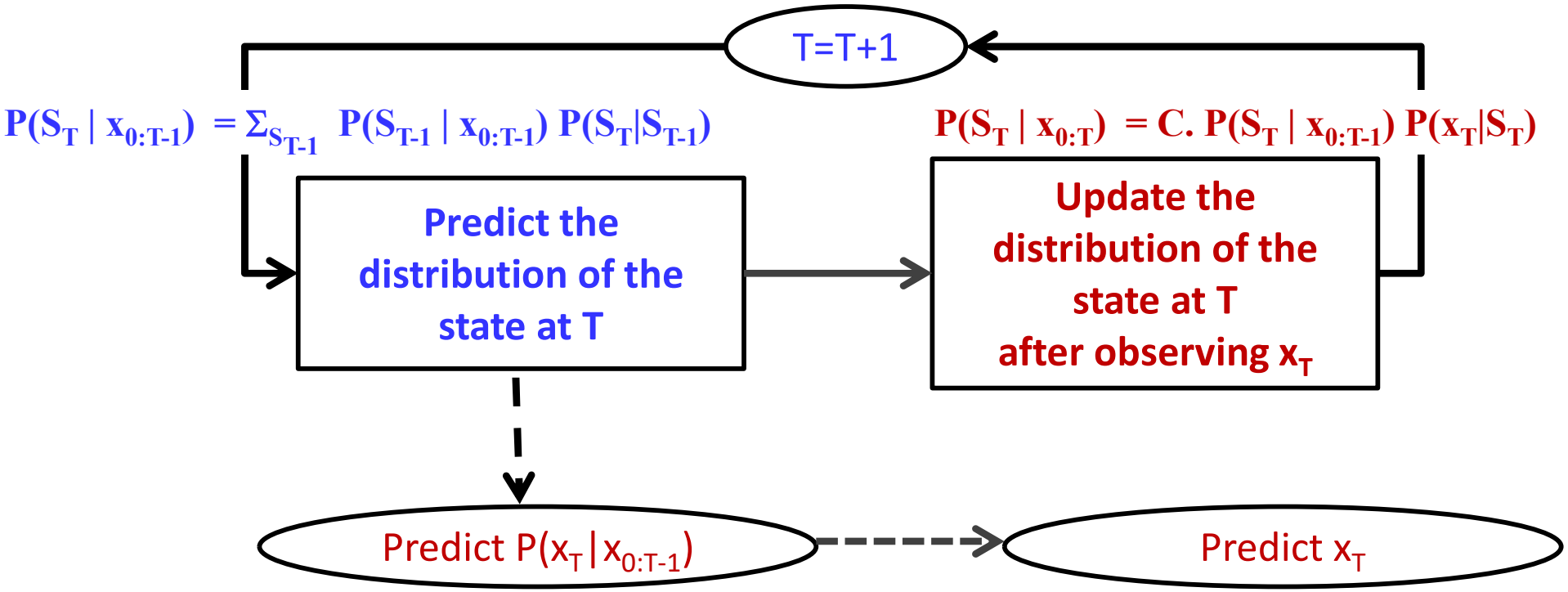
$$\boldsymbol{P(S|O) = N(\mu_S + \Theta_S A^{\mathrm{T}}\left(A\Theta_S A^{\mathrm{T}} + \Theta_\varepsilon\right)^{-1}(O - A\mu_s - \mu_\varepsilon),}$$
$$\boldsymbol{\Theta_S - \Theta_S A^{\mathrm{T}}\left(A\Theta_S A^{\mathrm{T}} + \Theta_\varepsilon\right)^{-1} A\Theta_S)}$$

Estimating the *state*



- The state is estimated from the updated distribution
  - The updated distribution is propagated into time, not the state

Predicting the *next observation*



- The probability distribution for the observations at the next time is a mixture:
- $P(X_t|X_{0:t-1}) = \sum_{S_t} P(X_t|S_t)P(S_t|X_{0:t-1})$
- The actual observation can be predicted from $P(x_T|x_{0:T-1})$

Predicting the next observation

- Can use any of the various estimators of x_T from $P(x_T|x_{0:T-1})$
- MAP estimate:
  - $\text{argmax}_{x_T} P(x_T|x_{0:T-1})$
- MMSE estimate:
  - $\text{Expectation}(x_T|x_{0:T-1})$

Difference from Viterbi decoding

- Estimating only the *current* state at any time
  - Not the state sequence
  - Although we are considering all past observations
- The most likely state at T and T+1 may be such that there is no valid transition between S_T and S_{T+1}

The real-valued state model

- A state equation describing the dynamics of the system

$$s_t = f(s_{t-1}, \varepsilon_t)$$

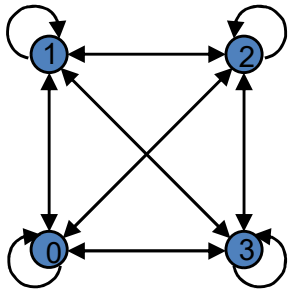
  - s_t is the state of the system at time t
  - ε_t is a driving function, which is assumed to be random
- The state of the system at any time depends only on the state at the previous time instant and the driving term at the current time

- An observation equation relating state to observation

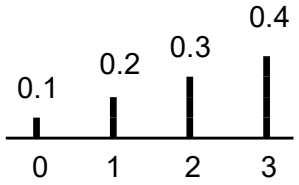
$$o_t = g(s_t, \gamma_t)$$

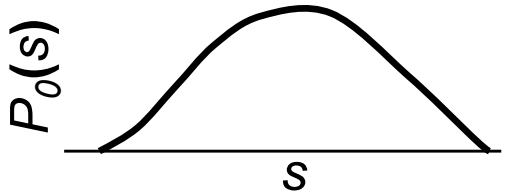
  - o_t is the observation at time t
  - γ_t is the noise affecting the observation (also random)
- The observation at any time depends only on the current state of the system and the noise

Discrete vs. Continuous state systems



$\pi =$





$$s_t = f(s_{t-1}, \varepsilon_t)$$

$$o_t = g(s_t, \gamma_t)$$

Prediction at time 0:

$$P(S_0) = \pi(S_0) \qquad P(S_0) = P_0(S_0)$$

Update after O_0:

$$P(S_0|O_0) = C.\pi(S_0)P(O_0|S_0) \qquad P(S_0|O_0) = C.P(S_0)P(O_0|S_0)$$

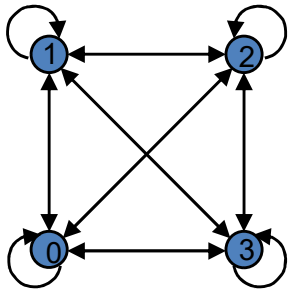
Prediction at time 1:

$$P(S_1|O_0) = \sum_{S_0} P(S_0|O_0)P(S_1|S_0) \qquad P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0)P(S_1|S_0)dS_0$$

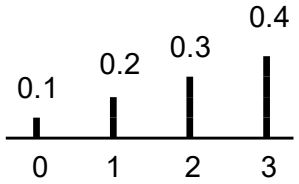
Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1) \qquad P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1)$$

Discrete vs. Continuous State Systems



$\pi =$



$$s_t = f(s_{t-1}, \varepsilon_t)$$

$$o_t = g(s_t, \gamma_t)$$

Prediction at time t:

$$P(S_t|O_{0:t-1}) = \sum_{S_{t-1}} P(S_{t-1}|O_{0:t-1})P(S_t|S_{t-1})$$

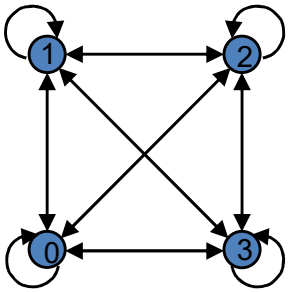
$$P(S_t|O_{0:t-1}) = \int_{-\infty}^{\infty} P(S_{t-1}|O_{0:t-1})P(S_t|S_{t-1})dS_{t-1}$$

Update after observing O_t:

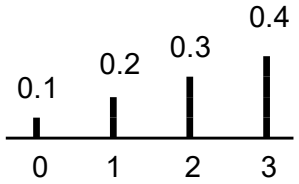
$$P(S_t|O_{0:t}) = C.P(S_t|O_{0:t-1})P(O_t|S_t)$$

$$P(S_t|O_{0:t}) = C.P(S_t|O_{0:t-1})P(O_t|S_t)$$

Discrete vs. Continuous State Systems



$\pi =$



$$s_t = f(s_{t-1}, \varepsilon_t)$$

$$o_t = g(s_t, \gamma_t)$$

Parameters		
Initial state prob.	π	$P(s)$
Transition prob	$P(s_t = j \mid s_{t-1} = i)$	$P(s_t \mid s_{t-1})$
Observation prob	$P(O \mid s)$	$P(O \mid s)$

Special case: Linear Gaussian model

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$P(\varepsilon) = \frac{1}{\sqrt{(2\pi)^d \,|\Theta_\varepsilon|}} \exp\left(-0.5(\varepsilon - \mu_\varepsilon)^T \Theta_\varepsilon^{-1}(\varepsilon - \mu_\varepsilon)\right)$$

$$o_t = B_t s_t + \gamma_t$$

$$P(\gamma) = \frac{1}{\sqrt{(2\pi)^d \,|\Theta_\gamma|}} \exp\left(-0.5(\gamma - \mu_\gamma)^T \Theta_\gamma^{-1}(\gamma - \mu_\gamma)\right)$$

- A *linear* state dynamics equation
  - Probability of state driving term ε is Gaussian
  - Sometimes viewed as a driving term μ_ε and additive zero-mean noise
- A *linear* observation equation
  - Probability of observation noise γ is Gaussian
- A_t, B_t and Gaussian parameters assumed known
  - May vary with time

Linear model example
The wind and the target

- **State:** Wind speed at time t depends on speed at time t-1

$$S_t = S_{t-1} + \epsilon_t$$

- **Observation:** Arrow position at time t depends on wind speed at time t

$$O_t = BS_t + \gamma_t$$

Model Parameters: The initial state probability

$$P_0(s) = \frac{1}{\sqrt{(2\pi)^d \mid R \mid}} \exp\left(-0.5(s-\bar{s})R^{-1}(s-\bar{s})^T\right)$$

$$P_0(s) = Gaussian(s; \bar{s}, R)$$

- We also assume the *initial* state distribution to be Gaussian
  - Often assumed zero mean

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Model Parameters: The observation probability

$$o_t = B_t s_t + \gamma_t \qquad P(\gamma) = Gaussian(\gamma; \mu_\gamma, \Theta_\gamma)$$

$$P(o_t \mid s_t) = Gaussian(o_t; \mu_\gamma + B_t s_t, \Theta_\gamma)$$

- The probability of the observation, given the state, is simply the probability of the noise, with the mean shifted
  - Since the only uncertainty is from the noise
- The new mean is the mean of the distribution of the noise + the value of the observation in the absence of noise

Model Parameters: State transition probability

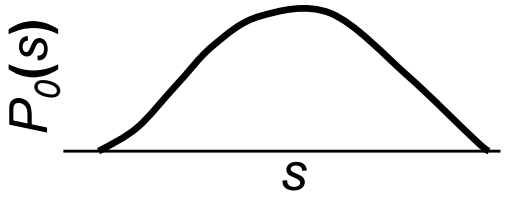
$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$P(\varepsilon) = Gaussian(\varepsilon; \mu_\varepsilon, \Theta_\varepsilon)$$

$$P(s_{t+1} \mid s_t) = Gaussian(s_t; \mu_\varepsilon + A_t s_t, \Theta_\varepsilon)$$

- The probability of the state at time t, given the state at t-1, is simply the probability of the driving term, with the mean shifted

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$P(S_0) = P_0(S_0)$$

Update after O_0:

$$P(S_0|O_0) = C.P(S_0)P(O_0|S_0)$$

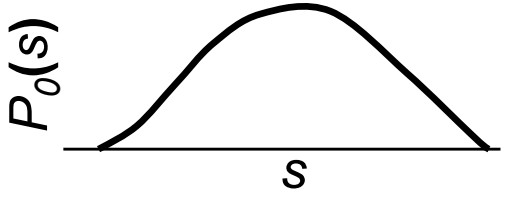
Prediction at time 1:

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0)P(S_1|S_0)dS_0$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1)$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$P(S_0) = P_0(S_0)$$

Update after O_0:

$$P(S_0|O_0) = C.P(S_0)P(O_0|S_0)$$

Prediction at time 1:

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0)P(S_1|S_0)dS_0$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1)$$

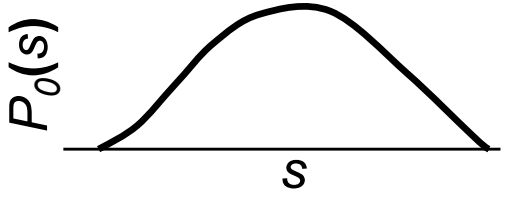
Model Parameters: The initial state probability

$$P_0(s) = \frac{1}{\sqrt{(2\pi)^d \mid R_0 \mid}} \exp\left(-0.5(s-\bar{s}_0)R_0^{-1}(s-\bar{s}_0)^T\right)$$

$$P_0(s) = Gaussian(s; \bar{s}_0, R_0)$$

- We assume the *initial* state distribution to be Gaussian
  - Often assumed zero mean

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

a priori probability distribution of state s

Prediction at time 0:

$$P(S_0) = P_0(S_0)$$

$$= N(\bar{s}_0, R_0)$$

Update after O_0:

$$P(S_0|O_0) = C.P(S_0)P(O_0|S_0)$$

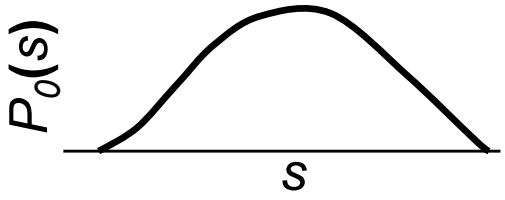
Prediction at time 1:

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0)P(S_1|S_0)dS_0$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1)$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = C.P(S_0)P(O_0|S_0)$$

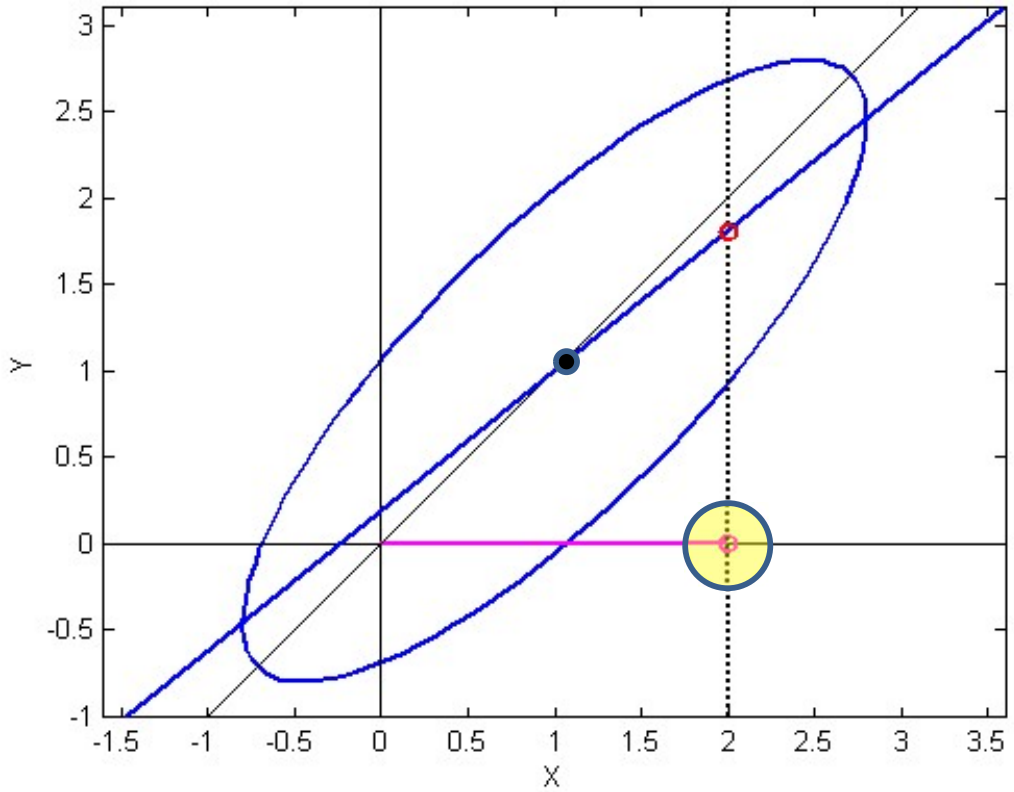
Prediction at time 1:

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0)P(S_1|S_0)dS_0$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1)$$

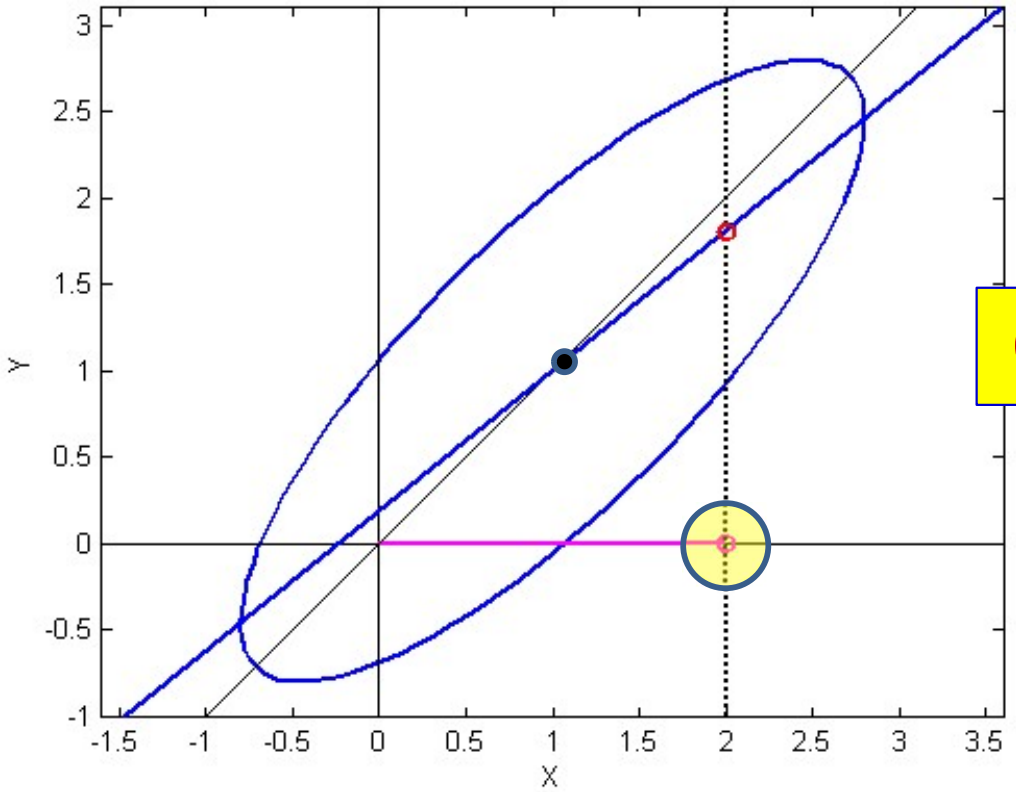
Recap: Conditional of S given O: P(S|O) for Gaussian RVs



$$O = BS + \gamma$$

$$P(S|O) = N(\mu_S + \Theta_{SO}\Theta_O^{-1}(O - \mu_O),\ \Theta_S - \Theta_{SO}\Theta_O^{-1}\Theta_{OS})$$

Recap: Conditional of S given O: P(S|O) for Gaussian RVs



$$O = BS + \gamma$$

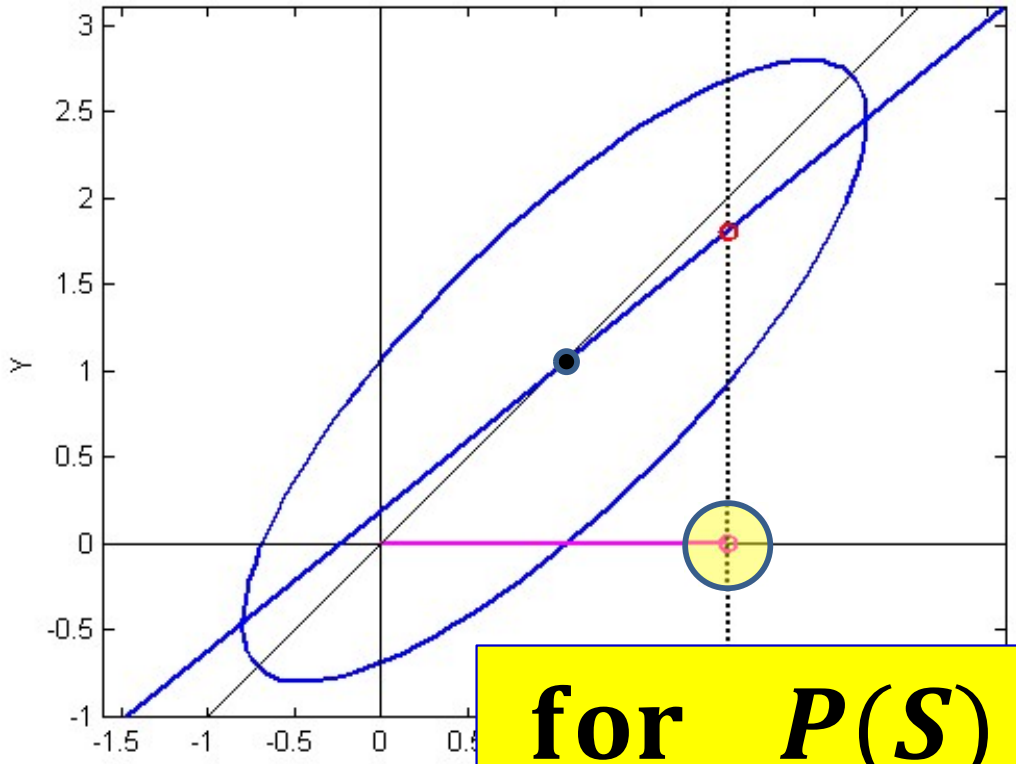
$$\Theta_{SO} = \Theta_S B^{\mathrm{T}}$$

$$\Theta_O = B\Theta_S B^{\mathrm{T}} + \Theta_\gamma$$

$$P(S|O) = N(\mu_S + \Theta_{SO}\Theta_O^{-1}(O - \mu_O),\ \Theta_S - \Theta_{SO}\Theta_O^{-1}\Theta_{OS})$$

$$P(S|O) = N(\mu_S + \Theta_S B^{\mathrm{T}}\left(B\Theta_S B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}(O - B\mu_s - \mu_\gamma),$$
$$\Theta_S - \Theta_S B^{\mathrm{T}}\left(B\Theta_S B^{\mathrm{T}} + \Theta_\gamma\right)^{-1} B\Theta_S)$$

Recap: Conditional of S given O: P(S|O) for Gaussian RVs



$$O = BS + \gamma$$

for $P(S) = N(\bar{s}_0, R_0)$

$$P(S_0|O_0) = N(\bar{s_0} + R_0 B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}(O_0 - B\bar{s}_0 - \mu_\gamma),$$
$$R_0 - R_0 B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1} BR_0)$$

Recap: Conditional of S given O: P(S|O) for Gaussian RVs

$$O = BS + \gamma$$

$$K_0 = R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma \right)^{-1}$$

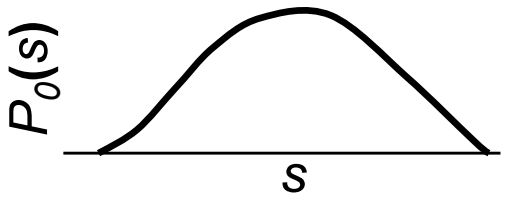
$$\hat{s}_0 = \bar{s}_0 + K_0 (O_0 - B\bar{s}_0 - \mu_\gamma)$$

$$\hat{R}_0 = (I - K_0)\, R_0$$

$$P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)$$

$$P(S_0|O_0) = N(\bar{s}_0 + R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma \right)^{-1} (O_0 - B\bar{s}_0 - \mu_\gamma),$$
$$R_0 - R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma \right)^{-1} B R_0)$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = C.P(S_0)P(O_0|S_0)$$

$$P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)$$

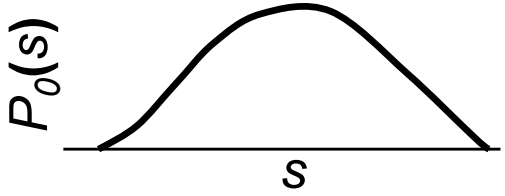
Prediction at time 1:

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0)P(S_1|S_0)dS_0$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1)$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)$$

$$\boldsymbol{K_0 = R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_0 = \bar{s}_0 + K_0 (O_0 - B\bar{s}_0 - \mu_\gamma)}$$

$$\boldsymbol{\hat{R}_0 = (I - K_0)\, R_0}$$

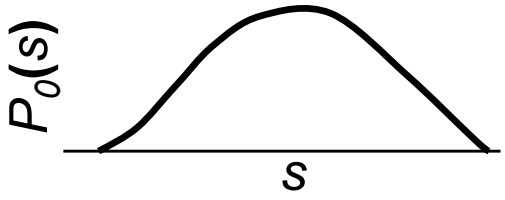
Prediction at time 1:

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0) P(S_1|S_0) dS_0$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0) P(O_1|S_1)$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = C.P(S_0)P(O_0|S_0)$$

$$= N(\bar{s}_0 + R_0 B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}(O_0 - B\bar{s}_0 - \mu_\gamma),$$
$$R_0 - R_0 B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1} BR_0)$$

Prediction at time 1:

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0)P(S_1|S_0)dS_0$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1)$$

Introducting shorthand notation

$$P(S_0|O_0) = N(\bar{s}_0 + R_0 B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}(O_0 - B\bar{s}_0 - \mu_\gamma),$$
$$R_0 - R_0 B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1} BR_0)$$

$$\hat{\boldsymbol{s}}_0 = \bar{\boldsymbol{s}}_0 + \boldsymbol{R}_0\boldsymbol{B}^{\mathrm{T}}\left(\boldsymbol{B}\boldsymbol{R}_0\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{\Theta}_\gamma\right)^{-1}(\boldsymbol{O} - \boldsymbol{B}\bar{\boldsymbol{s}}_0 - \boldsymbol{\mu}_\gamma)$$

$$\hat{\boldsymbol{R}}_0 = \boldsymbol{R}_0 - \boldsymbol{R}_0\boldsymbol{B}^{\mathrm{T}}\left(\boldsymbol{B}\boldsymbol{R}_0\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{\Theta}_\gamma\right)^{-1}\boldsymbol{B}\boldsymbol{R}_0$$

$$\boldsymbol{P}(\boldsymbol{S}_0|\boldsymbol{O}_0) = \boldsymbol{N}(\hat{\boldsymbol{s}}_0, \hat{\boldsymbol{R}}_0)$$

Introducting shorthand notation

$$P(S_0|O_0) = N(\bar{s}_0 + R_0 B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}(O_0 - B\bar{s}_0 - \mu_\gamma),$$
$$R_0 - R_0 B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1} BR_0)$$

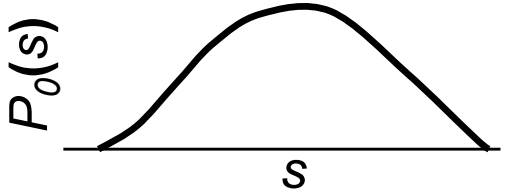
$$\boldsymbol{K_0 = R_0 B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_0 = \bar{s}_0 + K_0\left(O - B\bar{s}_0 - \mu_\gamma\right)}$$

$$\boldsymbol{\hat{R}_0 = (I - K_0 B)R_0}$$

$$\boldsymbol{P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)}$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)$$

$$\boldsymbol{K_0 = R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_0 = \bar{s}_0 + K_0 (O_0 - B\bar{s}_0 - \mu_\gamma)}$$

$$\boldsymbol{\hat{R}_0 = (I - K_0)\, R_0}$$

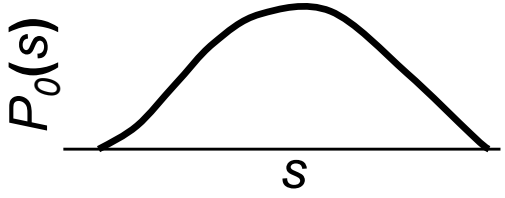
Prediction at time 1:

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0) P(S_1|S_0) dS_0$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0) P(O_1|S_1)$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)$$

$$\boldsymbol{K_0 = R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_0 = \bar{s}_0 + K_0(O_0 - B\bar{s}_0 - \mu_\gamma)}$$

$$\boldsymbol{\hat{R}_0 = (I - K_0)\, R_0}$$

Prediction at time 1:

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0) P(S_1|S_0) dS_0$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.\, P(S_1|O_0) P(O_1|S_1)$$

The prediction equation

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0)P(S_1|S_0)dS_0$$

$$P(S_0|O_0) = N(\hat{\boldsymbol{s}}_0, \hat{\boldsymbol{R}}_0)$$

$$P(S_1|S_0) = N(\boldsymbol{A}S_0 + \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}, \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}})$$

$$P(\varepsilon) = N(\mu_\varepsilon, \Theta_\varepsilon)$$

$$s_{t+1} = A_t s_t + \varepsilon_t$$

- The integral of the product of two Gaussians

$$P(S_1|O_0) = \int_{-\infty}^{\infty} Gaussian(S_0; \hat{s}_0, \hat{R}_0)Gaussian(S_1; AS_0, \Theta_\varepsilon)dS_0$$

The Prediction Equation

- The integral of the product of two Gaussians is Gaussian!

$$P(S_1|O_0) = \int_{-\infty}^{\infty} Gaussian(S_0; \hat{s}_0, \hat{R}_0) Gaussian(S_1; AS_0 + \mu_\varepsilon, \Theta_\varepsilon) dS_0$$

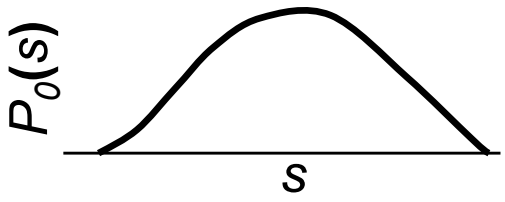
$$= \int_{-\infty}^{\infty} C_1 exp(-0.5(S_0 - \hat{s}_0)\hat{R}_0^{-1}(S_0 - \hat{s}_0)^T) . C_2 exp(-0.5(S_1 - AS_0 - \mu_\varepsilon)\Theta_\varepsilon^{-1}(S_1 - AS_0 - \mu_\varepsilon)^T) dS_0$$

$$= Gaussian(S_1; A\hat{s}_0 + \mu_\varepsilon, \Theta_\varepsilon + A\hat{R}_0 A^T)$$

$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$P(S_1|O_0) = N(A\hat{s}_0 + \mu_\varepsilon, \Theta_\varepsilon + A\hat{R}_0 A^T)$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)$$

$$\boldsymbol{K_0 = R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_0 = \bar{s}_0 + K_0 (O_0 - B\bar{s}_0 - \mu_\gamma)}$$

$$\boldsymbol{\hat{R}_0 = (I - K_0)\, R_0}$$

Prediction at time 1:

$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0) P(S_1|S_0) dS_0$$

$$= N(A\hat{s}_0 + \mu_\varepsilon, \Theta_\varepsilon + A\hat{R}_0 A^T)$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0) P(O_1|S_1)$$

More shorthand notation

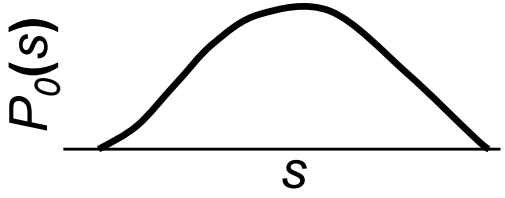
$$P(S_1|O_0) = N(A\hat{s}_0 + \mu_\varepsilon, \Theta_\varepsilon + A\hat{R}_0 A^T)$$

$$\bar{s}_1 = A\hat{s}_0 + \mu_\varepsilon$$

$$R_1 = \Theta_\varepsilon + A\hat{R}_0 A^T$$

$$P(S_1|O_0) = N(\bar{s}_1, R_1)$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)$$

$$\boldsymbol{K_0 = R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_0 = \bar{s}_0 + K_0(O_0 - B\bar{s}_0 - \mu_\gamma)}$$

$$\boldsymbol{\hat{R}_0 = (I - K_0)\, R_0}$$

Prediction at time 1:

$$P(S_1|O_0) = N(\bar{s}_1, R_1)$$

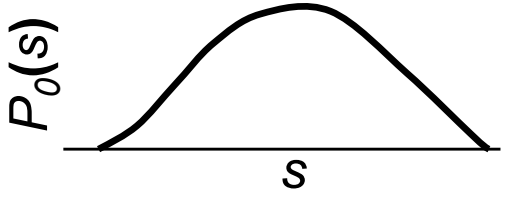
$$\boldsymbol{\bar{s}_1 = A\hat{s}_0 + \mu_\varepsilon}$$

$$\boldsymbol{R_1 = \Theta_\varepsilon + A\hat{R}_0 A^T}$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1)$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)$$

$$\boldsymbol{K_0 = R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_0 = \bar{s}_0 + K_0(O_0 - B\bar{s}_0 - \mu_\gamma)}$$

$$\boldsymbol{\hat{R}_0 = (I - K_0)\, R_0}$$

Prediction at time 1:

$$P(S_1|O_0) = N(\bar{s}_1, R_1)$$

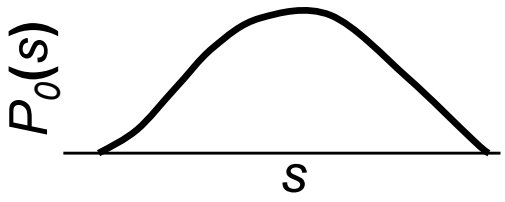
$$\boldsymbol{\bar{s}_1 = A\hat{s}_0 + \mu_\varepsilon}$$

$$\boldsymbol{R_1 = \Theta_\varepsilon + A\hat{R}_0 A^T}$$

Update after O_1:

$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1)$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)$$

$$\boldsymbol{K_0 = R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_0 = \bar{s}_0 + K_0(O_0 - B\bar{s}_0 - \mu_\gamma)}$$

$$\boldsymbol{\hat{R}_0 = (I - K_0 B)\, R_0}$$

Prediction at time 1:

$$P(S_1|O_0) = N(\bar{s}_1, R_1)$$

$$\boldsymbol{\bar{s}_1 = A\hat{s}_0 + \mu_\varepsilon}$$

$$\boldsymbol{R_1 = \Theta_\varepsilon + A\hat{R}_0 A^T}$$

Update after O_1:

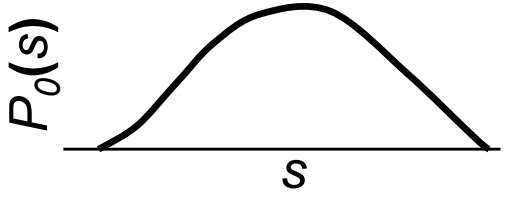
$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1) = N(\hat{s}_1, \hat{R}_1)$$

$$\boldsymbol{K_1 = R_1 B^{\mathrm{T}} \left(B R_1 B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_1 = \bar{s}_1 + K_1(O_1 - B\bar{s}_1 - \mu_\gamma)}$$

$$\boldsymbol{\hat{R}_1 = (I - K_1 B)\, R_1}$$

Continuous state systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

$$\boldsymbol{P(S_0) = N(\bar{s}_0, R_0)}$$

Update after O_0:

$$P(S_0|O_0) = N(\hat{s}_0, \hat{R}_0)$$

$$\boldsymbol{K_0 = R_0 B^{\mathrm{T}} \left(B R_0 B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_0 = \bar{s}_0 + K_0(O_0 - B\bar{s}_0 - \mu_\gamma)}$$

$$\boldsymbol{\hat{R}_0 = (I - K_0 B)\, R_0}$$

Prediction at time 1:

$$P(S_1|O_0) = N(\bar{s}_1, R_1)$$

$$\boldsymbol{\bar{s}_1 = A\hat{s}_0 + \mu_\varepsilon}$$

$$\boldsymbol{R_1 = \Theta_\varepsilon + A\hat{R}_0 A^T}$$

Update after O_1:

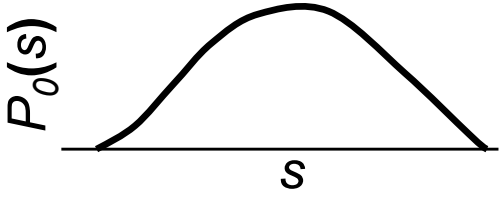
$$P(S_1|O_{0:1}) = N(\hat{s}_1, \hat{R}_1)$$

$$\boldsymbol{K_1 = R_1 B^{\mathrm{T}} \left(B R_1 B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

$$\boldsymbol{\hat{s}_1 = \bar{s}_1 + K_1(O_1 - B\bar{s}_1 - \mu_\gamma)}$$

$$\boldsymbol{\hat{R}_1 = (I - K_1 B)\, R_1}$$

Gaussian Continuous State Linear Systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

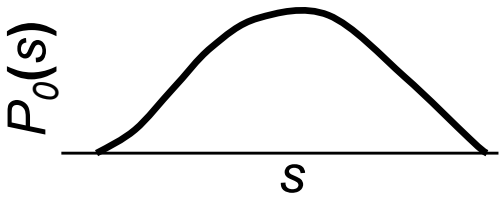
Prediction at time t:

$$P(S_t|O_{0:t-1}) = \int_{-\infty}^{\infty} P(S_{t-1}|O_{0:t-1}) P(S_t|S_{t-1}) dS_{t-1}$$

Update after observing O_t:

$$P(S_t|O_{0:t}) = C.P(S_t|O_{0:t-1}) P(O_t|S_t)$$

Gaussian Continuous State Linear Systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time t:

$$P(S_t|O_{0:t-1}) = N(\bar{s}_t, R_t)$$

$$\bar{s}_t = A\hat{s}_{t-1} + \mu_\varepsilon$$

$$R_t = \Theta_\varepsilon + A\hat{R}_{t-1}A^T$$

Update after observing O_t:

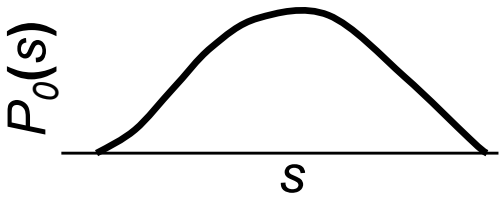
$$P(S_t|O_{0:t}) = N(\hat{s}_t, \hat{R}_t)$$

$$K_t = R_1 B^T \left(BR_1 B^T + \Theta_\gamma\right)^{-1}$$

$$\hat{s}_t = \bar{s}_t + Kt\,(Ot - B\bar{s}_t - \mu_\gamma)$$

$$\hat{R}_t = (I - KtB)\,R_t$$

Gaussian Continuous State Linear Systems



$$s_{t+1} = A_t s_t + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

Prediction at time t:

$$P(S_t|O_{0:t-1}) = N(\bar{s}_t, R_t)$$

Update after observing O_t:

$$P(S_t|O_{0:t}) = N(\hat{s}_t, \hat{R}_t)$$

KALMAN FILTER

$$\bar{s}_t = A\hat{s}_{t-1} + \mu_\varepsilon$$

$$R_t = \Theta_\varepsilon + A\hat{R}_{t-1}A^T$$

$$K_t = R_1 B^T \left(BR_1 B^T + \Theta_\gamma\right)^{-1}$$

$$\hat{s}_t = \bar{s}_t + Kt\,(Ot - B\bar{s}_t - \mu_\gamma)$$

$$\hat{R}_t = (I - KtB)\,R_t$$

The Kalman filter

- Prediction (based on state equation)

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon \qquad s_t = A_t s_{t-1} + \varepsilon_t$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

- Update (using observation and observation equation)

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma \right)^{-1} \qquad o_t = B_t s_t + \gamma_t$$

$$\hat{s}_t = \bar{s}_t + K_t \left(o_t - B_t \bar{s}_t - \mu_\gamma \right)$$

$$\hat{R}_t = \left(I - K_t B_t \right) R_t$$

Explaining the Kalman Filter

- Prediction

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

$$o_t = B_t s_t + \gamma_t$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

- The Kalman filter can be explained intuitively without working through the math

$$\hat{s}_t = \bar{s}_t + K_t (o_t - B_t \bar{s}_t - \mu_\gamma)$$

$$\hat{R}_t = (I - K_t B_t) R_t$$

The Kalman filter

- Prediction

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

The *predicted* state at time t is obtained simply by propagating the estimated state at t-1 through the state dynamics equation

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma \right)$$

$$\hat{s}_t = \bar{s}_t + K_t \left(o_t - B_t \bar{s}_t - \mu_\gamma \right)$$

$$\hat{R}_t = \left(I - K_t B_t \right) R_t$$

The Kalman filter

- Prediction

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

This is the uncertainty in the prediction. The variance of the predictor = variance of ε_t + variance of As_{t-1}

The two simply add because ε_t is not correlated with s_t

The Kalman filter

- Prediction

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

$$o_t = B_t s_t + \gamma_t$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

$$\hat{o}_t = B_t \bar{s}_t + \mu_\gamma$$

We can also predict the *observation* from the predicted state using the observation equation

$$\hat{s}_t = \bar{s}_t + K_t (o_t - B_t \bar{s}_t - \mu_\gamma)$$

$$\hat{R}_t = (I - K_t B_t) R_t$$

The Kalman filter

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

- Prediction

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

$$\hat{o}_t = B_t \bar{s}_t + \mu_\gamma$$

- Update

Actual observation

o_t

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma \right)^{-1}$$

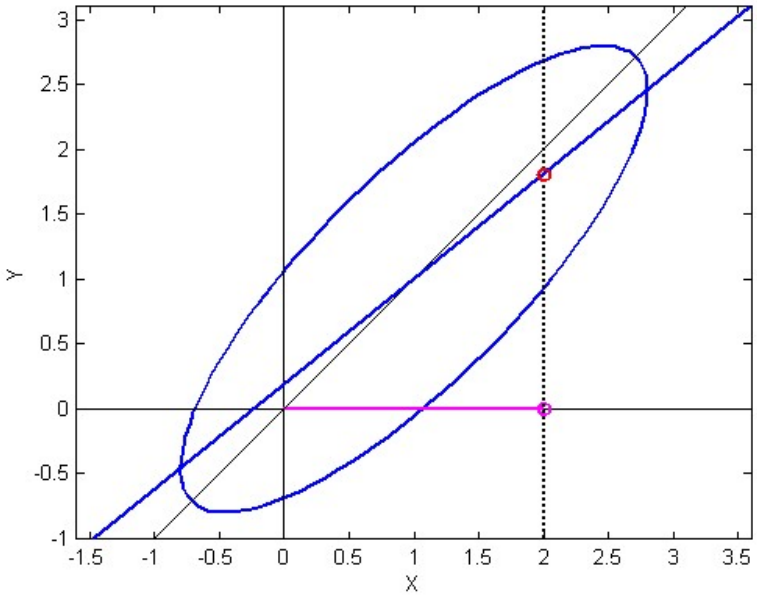
$$\hat{s}_t = \bar{s}_t + K_t \left(o_t - B_t \bar{s}_t \right)$$

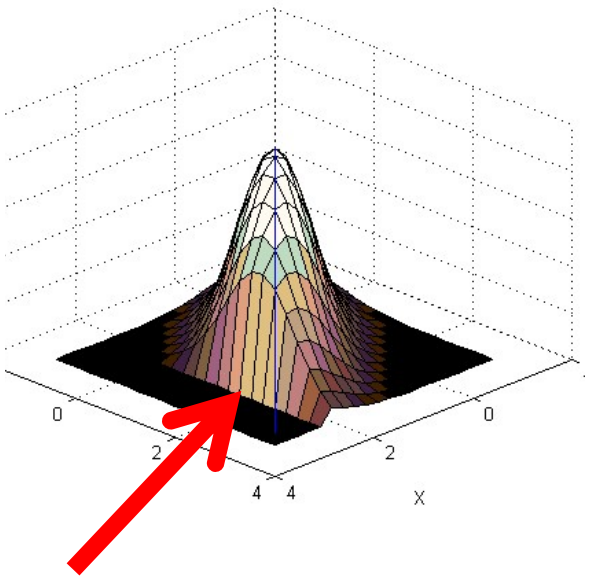
$$\hat{R}_t = \left(I - K_t B_t \right) R_t$$

MAP Recap (for Gaussians)

- If P(x,y) is Gaussian:

$$P(\mathbf{x},\mathbf{y}) = N(\begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}, \begin{bmatrix}C_{xx} & C_{xy} \\ C_{yx} & C_{yy}\end{bmatrix})$$





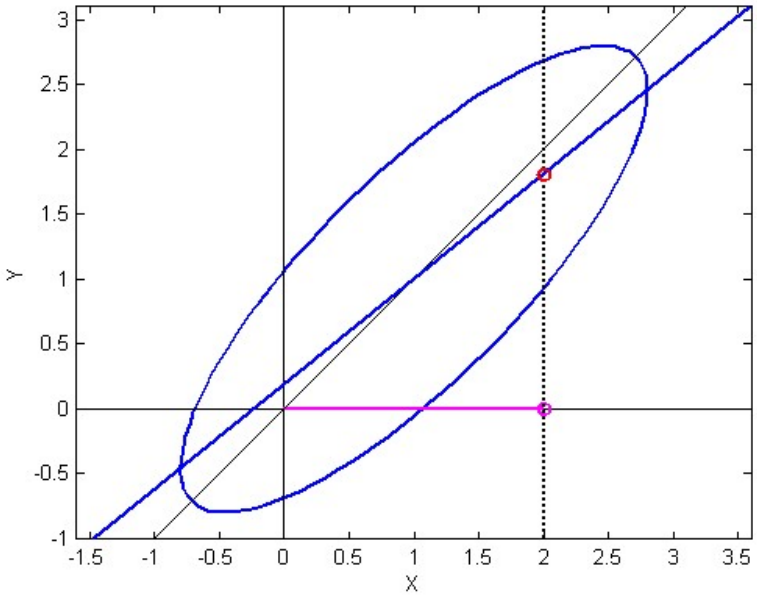
$$P(y \mid x) = N(\mu_y + C_{yx}C_{xx}^{-1}(x-\mu_x), C_{yy} - C_{yx}^{T}C_{xx}^{-1}C_{xy})$$

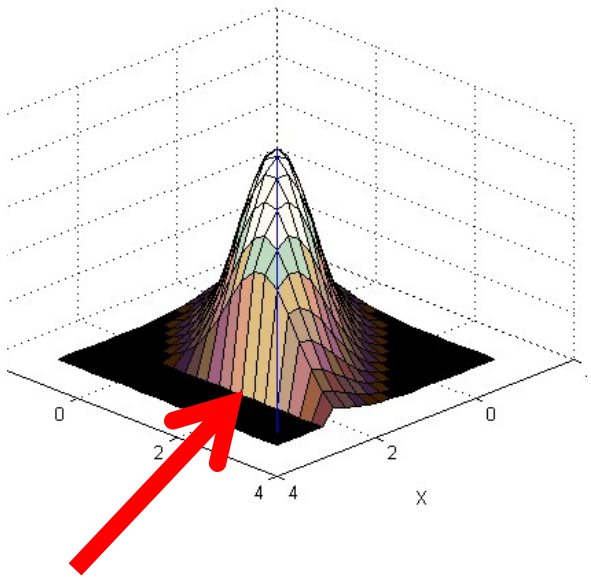
$$\hat{y} = \mu_y + C_{yx}C_{xx}^{-1}(x-\mu_x)$$

MAP Recap: For Gaussians

- If P(x,y) is Gaussian:

$$P(\mathbf{y},\mathbf{x}) = N\left(\begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix}, \begin{bmatrix}C_{xx} & C_{xy}\\ C_{yx} & C_{yy}\end{bmatrix}\right)$$





$$P(y \mid x) = N(\mu_y + C_{yx}C_{xx}^{-1}(x-\mu_x), C_{yy} - C_{yx}^{T}C_{xx}^{-1}C_{xy})$$

$$\hat{y} = \mu_y + C_{yx}C_{xx}^{-1}(x-\mu_x)$$

"Slope" of the line

The Kalman filter

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

- Prediction

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

$$\hat{o}_t = B_t \bar{s}_t + \mu_\gamma$$

o_t

- Update

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma \right)^{-1}$$

This is the slope of the MAP estimator that predicts s from o

$RB^T = C_{so}$, $(BRB^T+\Theta) = C_{oo}$

This is also called the **Kalman Gain**

The Kalman filter

- Prediction

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

We must correct the predicted value of the state after making an observation

$$\hat{o}_t = B_t \bar{s}_t + \mu_\gamma$$

o_t

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma \right)^{-1}$$

$$\hat{s}_t = \bar{s}_t + K_t \left(o_t - \hat{o}_t \right)$$

The correction is the difference between the *actual* observation and the *predicted* observation, scaled by the Kalman Gain

The Kalman filter

- Prediction

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

We must correct the predicted value of the state after making an observation

$$\hat{o}_t = B_t \bar{s}_t + \mu_\gamma$$

o_t

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma \right)^{-1}$$

$$\hat{s}_t = \bar{s}_t + K_t \left(o_t - B_t \bar{s}_t - \mu_\gamma \right)$$

The correction is the difference between the *actual* observation and the *predicted* observation, scaled by the Kalman Gain

The Kalman filter

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

- Prediction

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

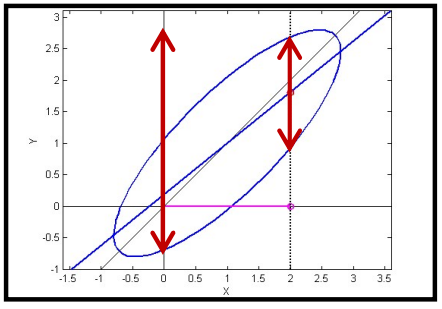
$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

- Update:

The uncertainty in state decreases if we observe the data and make a correction

The reduction is a multiplicative "shrinkage" based on Kalman gain and B

$$\hat{R}_t = (I - K_t B_t) R_t$$



The Kalman filter

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

- Prediction

$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

- Update:

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma \right)^{-1}$$

- Update

$$\hat{s}_t = \bar{s}_t + K_t \left(o_t - B_t \bar{s}_t - \mu_\gamma \right)$$

$$\hat{R}_t = \left(I - K_t B_t \right) R_t$$

The Kalman Filter

- Very popular for tracking the state of processes
  - Control systems
  - Robotic tracking
    - Simultaneous localization and mapping
  - Radars
  - Even the stock market..
- What are the parameters of the process?

Kalman filter contd.

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

- Model parameters A and B must be known
  - Often the state equation includes an *additional* driving term: $s_t = A_t s_{t-1} + G_t u_t + \varepsilon_t$
  - The parameters of the driving term must be known
- The initial state distribution must be known

Defining the parameters

- State state must be carefully defined
  - E.g. for a robotic vehicle, the state is an extended vector that includes the current velocity and acceleration
    - $S = [X, dX, d^2X]$
- State equation: Must incorporate appropriate constraints
  - If state includes acceleration and velocity, velocity at next time = current velocity + acc. * time step
  - $St = AS_{t-1} + e$
    - $A = [1\ t\ 0.5t^2;\ 0\ 1\ t;\ 0\ 0\ 1]$

Parameters

- Observation equation:
  - Critical to have accurate observation equation
  - Must provide a valid relationship between state and observations
- Observations typically high-dimensional
  - May have higher or lower dimensionality than state

Problems

$$s_t = f(s_{t-1}, \varepsilon_t)$$

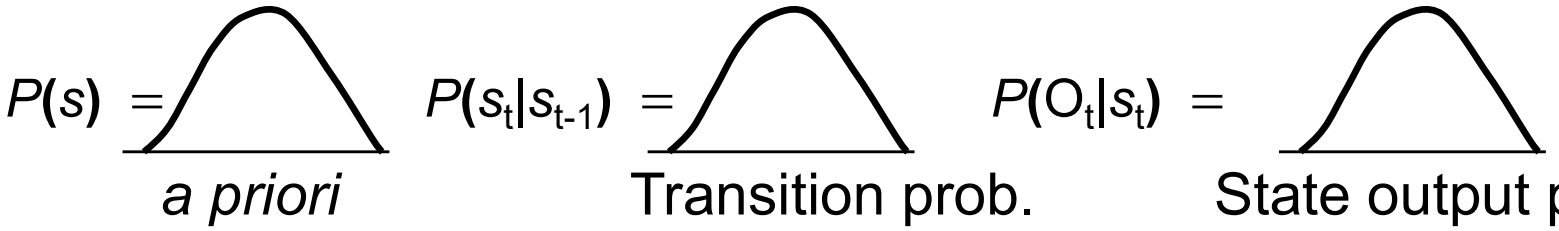
$$o_t = g(s_t, \gamma_t)$$

- f() and/or g() may not be nice linear functions
  - Conventional Kalman update rules are no longer valid
- ε and/or γ may not be Gaussian
  - Gaussian based update rules no longer valid

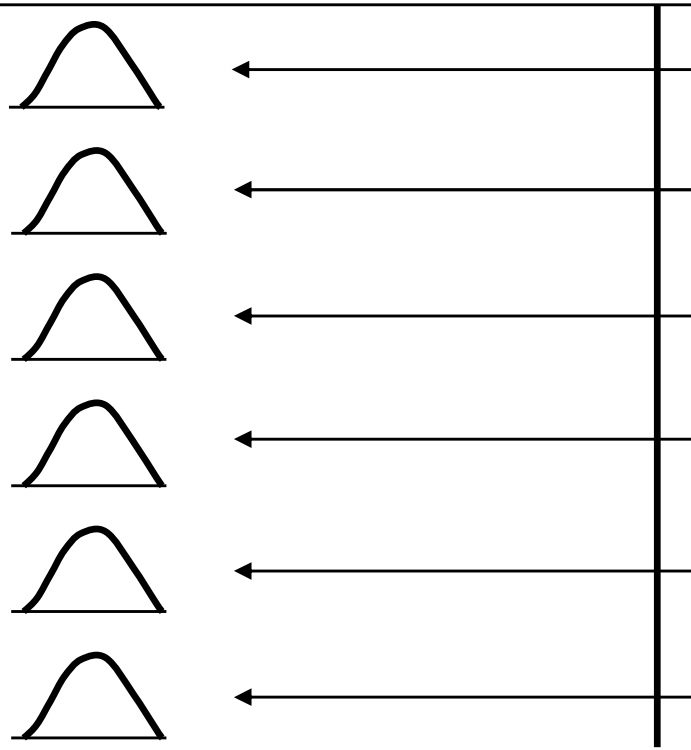
Linear Gaussian Model

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$



a priori Transition prob. State output prob



$P(s_0) = P(s)$

$P(s_0 | O_0) = C\, P(s_0)\, P(O_0 | s_0)$

$$P(s_1 | O_0) = \int_{-\infty}^{\infty} P(s_0 | O_0) P(s_1 | s_0) ds_0$$

$P(s_1 | O_{0:1}) = C\, P(s_1 | O_0)\, P(O_1 | s_0)$

$$P(s_2 | O_{0:1}) = \int_{-\infty}^{\infty} P(s_1 | O_{0:1}) P(s_2 | s_1) ds_1$$

$P(s_2 | O_{0:2}) = C\, P(s_2 | O_{0:1})\, P(O_2 | s_2)$

All distributions remain Gaussian

Problems

$$s_t = f(s_{t-1}, \varepsilon_t)$$

$$o_t = g(s_t, \gamma_t)$$

- Nonlinear f() and/or g() : The Gaussian assumption breaks down
  - Conventional Kalman update rules are no longer valid

The problem with non-linear functions

$$s_t = f(s_{t-1}, \varepsilon_t)$$

$$o_t = g(s_t, \gamma_t)$$

$$P(s_t \mid o_{0:t-1}) = \int_{-\infty}^{\infty} P(s_{t-1} \mid o_{0:t-1}) P(s_t \mid s_{t-1}) ds_{t-1}$$

$$P(s_t \mid o_{0:t}) = CP(s_t \mid o_{0:t-1}) P(o_t \mid s_t)$$

- Estimation requires knowledge of $P(o|s)$
  - Difficult to estimate for nonlinear $g()$
  - Even if it can be estimated, may not be tractable with update loop
- Estimation also requires knowledge of $P(s_t|s_{t-1})$
  - Difficult for nonlinear $f()$
  - May not be amenable to closed form integration

The problem with nonlinearity

$$o_t = g(s_t, \gamma_t)$$

- The PDF may not have a closed form

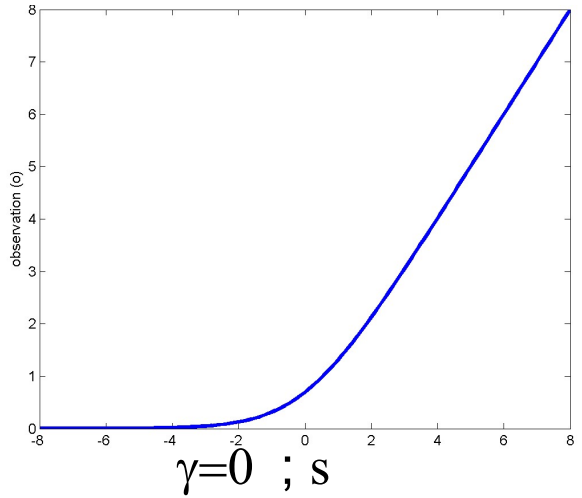
$$P(o_t \mid s_t) = \sum_{\gamma : g(s_t,\gamma)=o_t} \frac{P_\gamma(\gamma)}{|J_{g(s_t,\gamma)}(o_t)|}$$

$$|J_{g(s_t,\gamma)}(o_t)| = \begin{vmatrix} \frac{\partial o_t(1)}{\partial \gamma(1)} & \cdots & \frac{\partial o_t(1)}{\partial \gamma(n)} \\ \vdots & \ddots & \vdots \\ \frac{\partial o_t(n)}{\partial \gamma(1)} & \cdots & \frac{\partial o_t(n)}{\partial \gamma(n)} \end{vmatrix}$$

- Even if a closed form exists initially, it will typically become intractable very quickly

Example: a simple nonlinearity

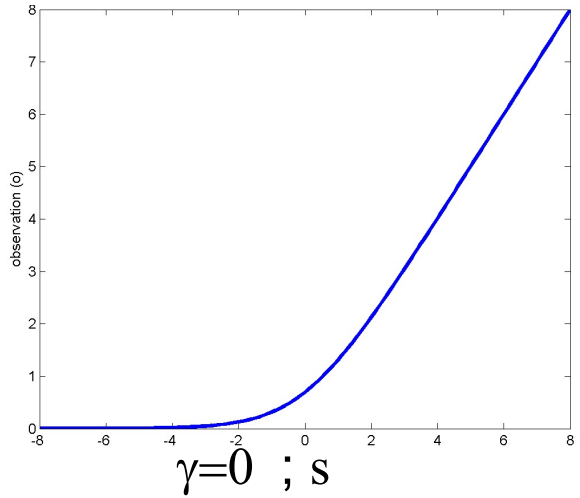
$$o = \gamma + \log(1 + \exp(s))$$



- P(o|s) = ?
  - Assume γ is Gaussian
  - $P(\gamma) = Gaussian(\gamma; \mu_\gamma, \Theta_\gamma)$

Example: a simple nonlinearity

$$o = \gamma + \log(1 + \exp(s))$$



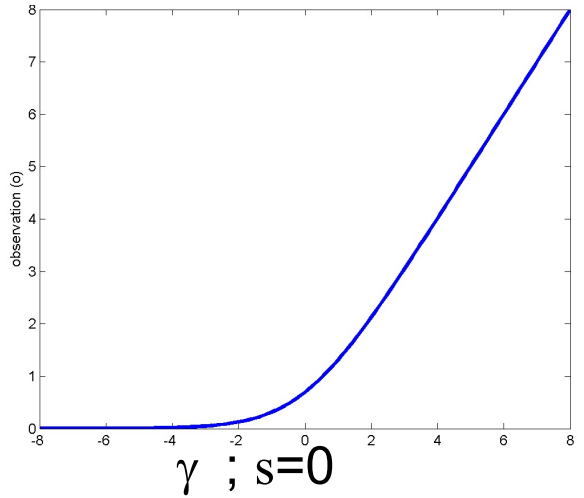
- P(o|s) = ?

$$P(\gamma) = Gaussian(\gamma; \mu_\gamma, \Theta_\gamma)$$

$$P(o \mid s) = Gaussian(o; \mu_\gamma + \log(1 + \exp(s)), \Theta_\gamma)$$

Example: At T=0.

$$o = \gamma + \log(1 + \exp(s))$$



- Assume initial probability P(s) is Gaussian

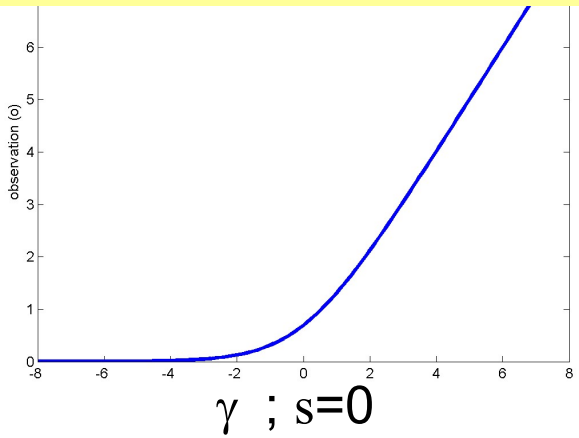
$$P(s_0) = P_0(s) = Gaussian(s; \bar{s}, R)$$

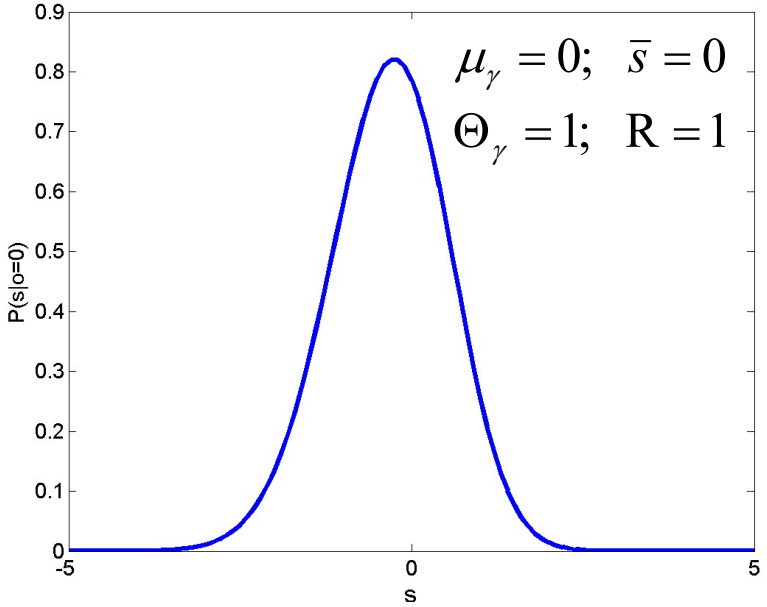
- Update $P(s_0 \mid o_0) = CP(o_0 \mid s_0)P(s_0)$

$$P(s_0 \mid o_0) = CGaussian(o; \mu_\gamma + \log(1 + \exp(s_0)), \Theta_\gamma)Gaussian(s_0; \bar{s}, R)$$

UPDATE: At T=0.

$$o = \gamma + \log(1 + \exp(s))$$





$$P(s_0 \mid o_0) = CGaussian(o; \mu_\gamma + \log(1 + \exp(s_0)), \Theta_\gamma) Gaussian(s_0; \bar{s}, R)$$

$$P(s_0 \mid o_0) = C \exp\left(\begin{array}{c} -0.5(\mu_\gamma + \log(1 + \exp(s_0)) - o)^T \Theta_\gamma^{-1} (\mu_\gamma + \log(1 + \exp(s_0)) - o) \\ -0.5(s_0 - \bar{s})^T R^{-1} (s_0 - \bar{s}) \end{array} \right)$$

- = Not Gaussian

Prediction for T = 1

$$s_t = s_{t-1} + \varepsilon \qquad P(\varepsilon) = Gaussian(\varepsilon; 0, \Theta_\varepsilon)$$

- Trivial, linear state transition equation

$$P(s_t \mid s_{t-1}) = Gaussian(s_t; s_{t-1}, \Theta_\varepsilon)$$

- Prediction $P(s_1 \mid o_0) = \int_{-\infty}^{\infty} P(s_0 \mid o_0) P(s_1 \mid s_0) ds_0$

$$P(s_1 \mid o_0) = \int_{-\infty}^{\infty} C \exp\left(\begin{array}{c} -0.5(\mu_\gamma + \log(1+\exp(s_0)) - o)^T \Theta_\gamma^{-1} (\mu_\gamma + \log(1+\exp(s_0)) - o) \\ -0.5(s_0 - \bar{s})^T R^{-1} (s_0 - \bar{s}) \end{array} \right) \exp\left((s_1 - s_0)^T \Theta_\varepsilon^{-1} (s_1 - s_0) \right) ds_0$$

- = intractable

Update at T=1 and later

- Update at T=1

$$P(s_t \mid o_{0:t}) = CP(s_t \mid o_{0:t-1})P(o_t \mid s_t)$$

  - Intractable

- Prediction for T=2

$$P(s_t \mid o_{0:t-1}) = \int_{-\infty}^{\infty} P(s_{t-1} \mid o_{0:t-1})P(s_t \mid s_{t-1})ds_{t-1}$$

  - Intractable

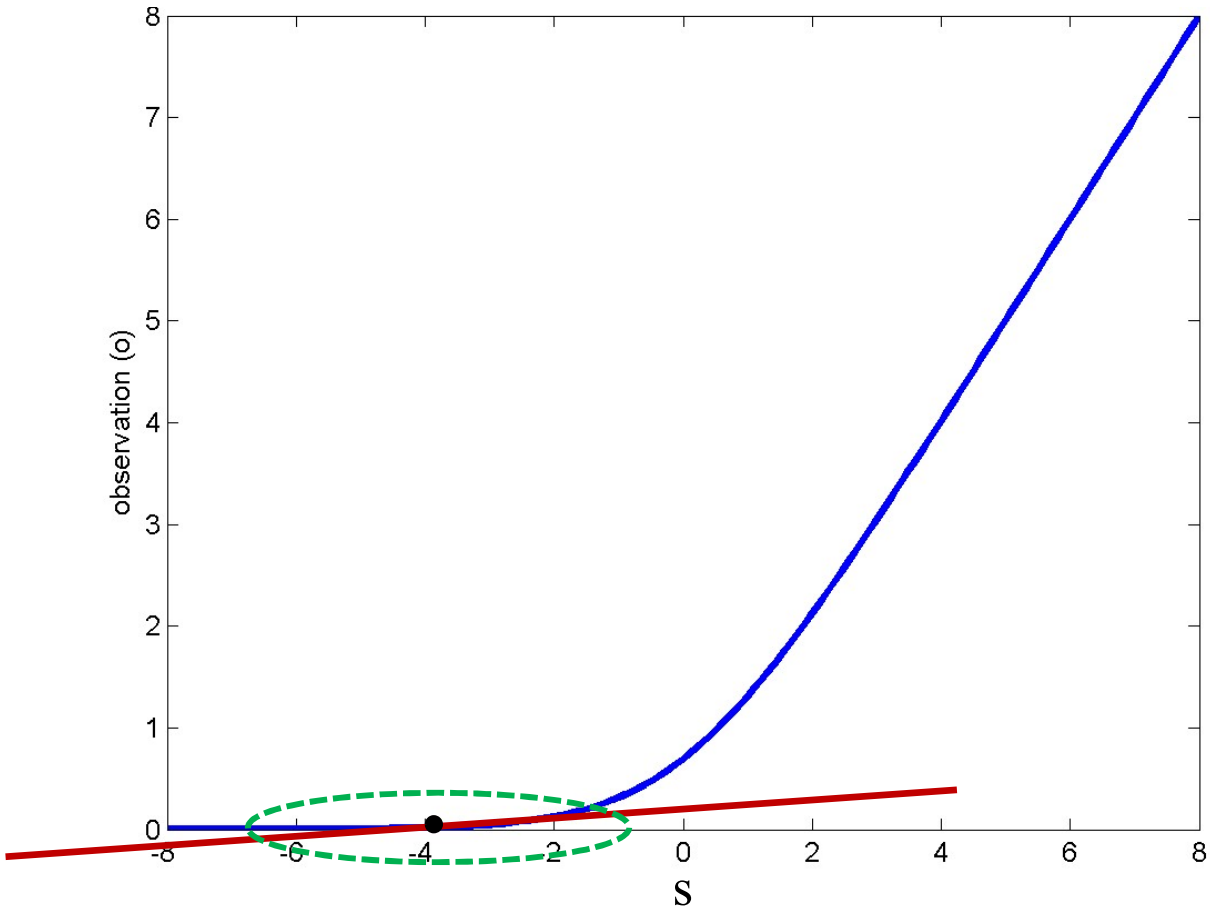
The State prediction Equation

$$s_t = f(s_{t-1}, \varepsilon_t)$$

- Similar problems arise for the state prediction equation

- $P(s_t|s_{t-1})$ may not have a closed form
- Even if it does, it may become intractable within the prediction and update equations
  - Particularly the prediction equation, which includes an integration operation

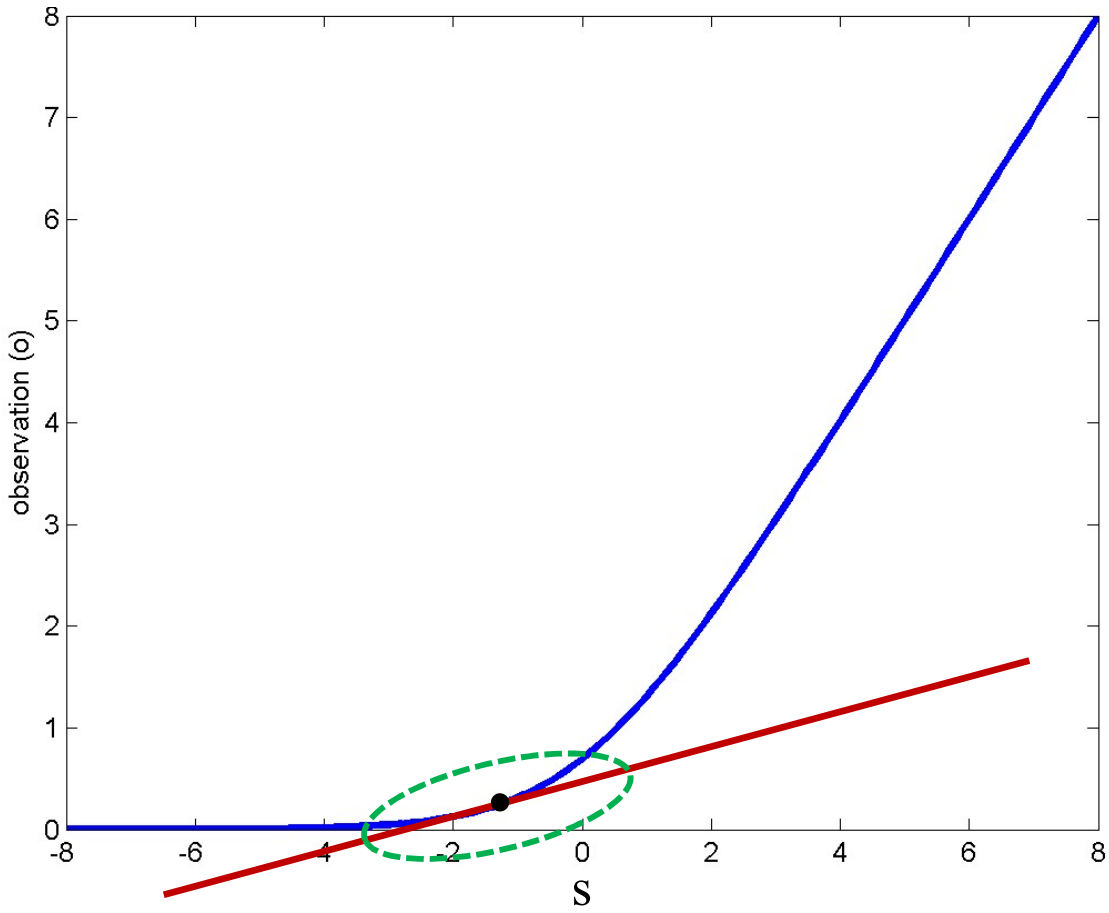
Simplifying the problem: Linearize



$$o = \gamma + \log(1 + \exp(s))$$

- The *tangent* at any point is a good *local* approximation if the function is sufficiently smooth

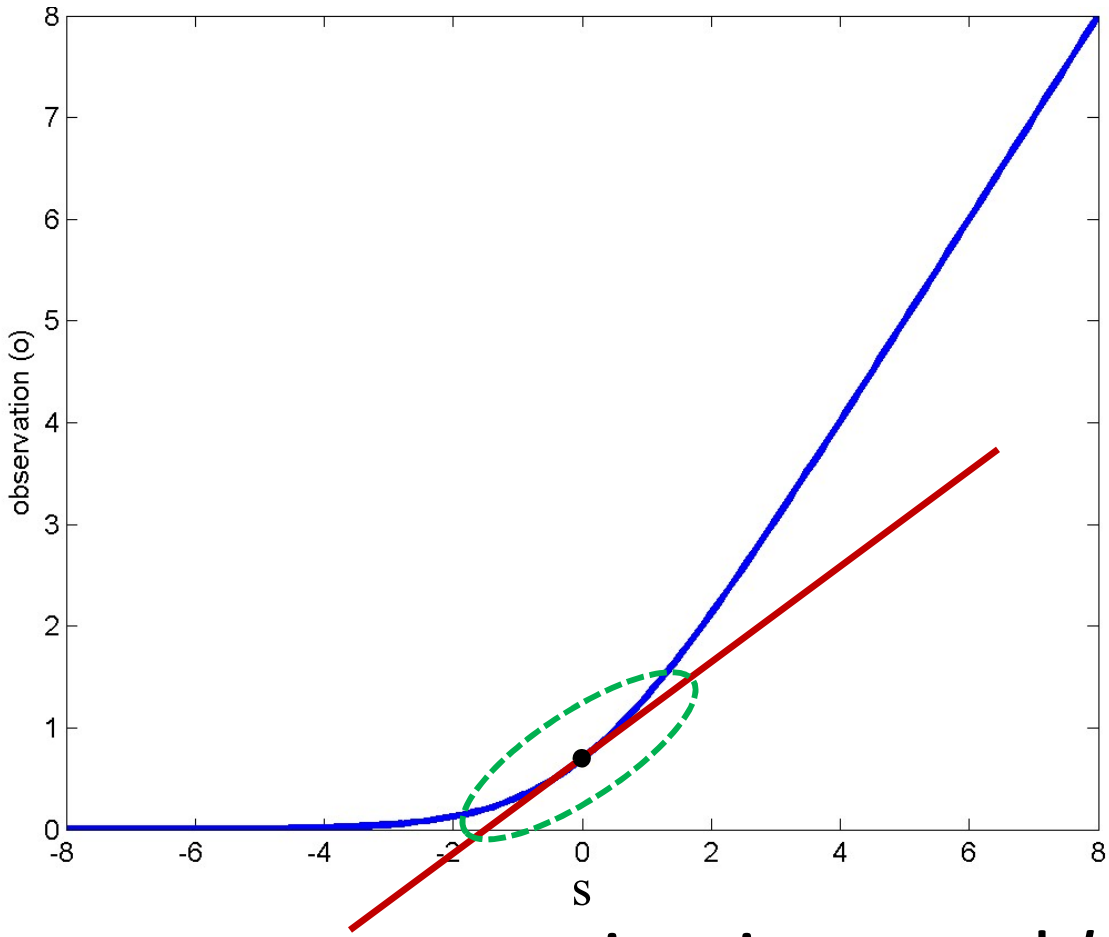
Simplifying the problem: Linearize



$$o = \gamma + \log(1 + \exp(s))$$

- The *tangent* at any point is a good *local* approximation if the function is sufficiently smooth

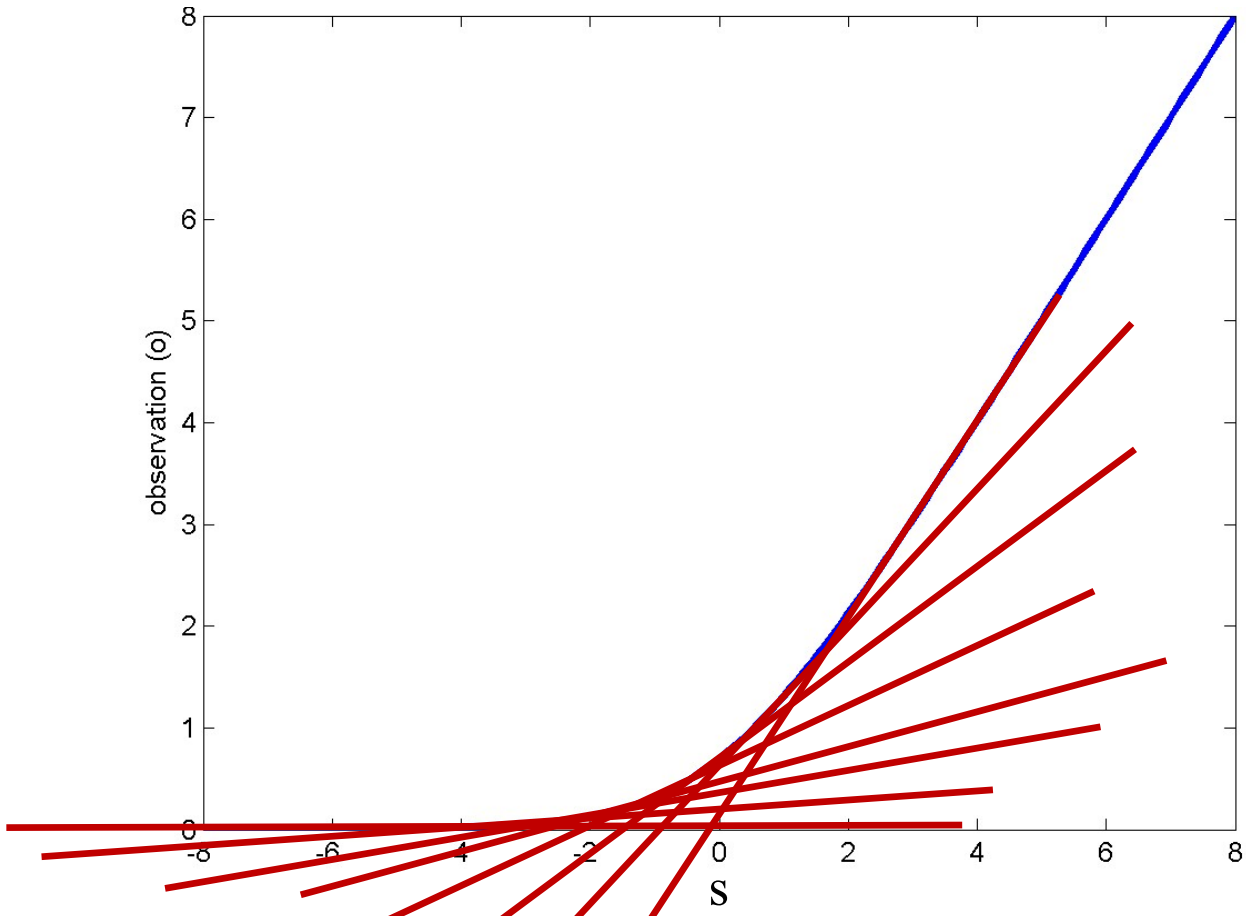
Simplifying the problem: Linearize



$$o = \gamma + \log(1 + \exp(s))$$

- The *tangent* at any point is a good *local* approximation if the function is sufficiently smooth

Simplifying the problem: Linearize



- The *tangent* at any point is a good *local* approximation if the function is sufficiently smooth

Linearizing the observation function

$$P(s_t \mid o_{0:t-1}) = Gaussian(\bar{s}_t, R_t)$$

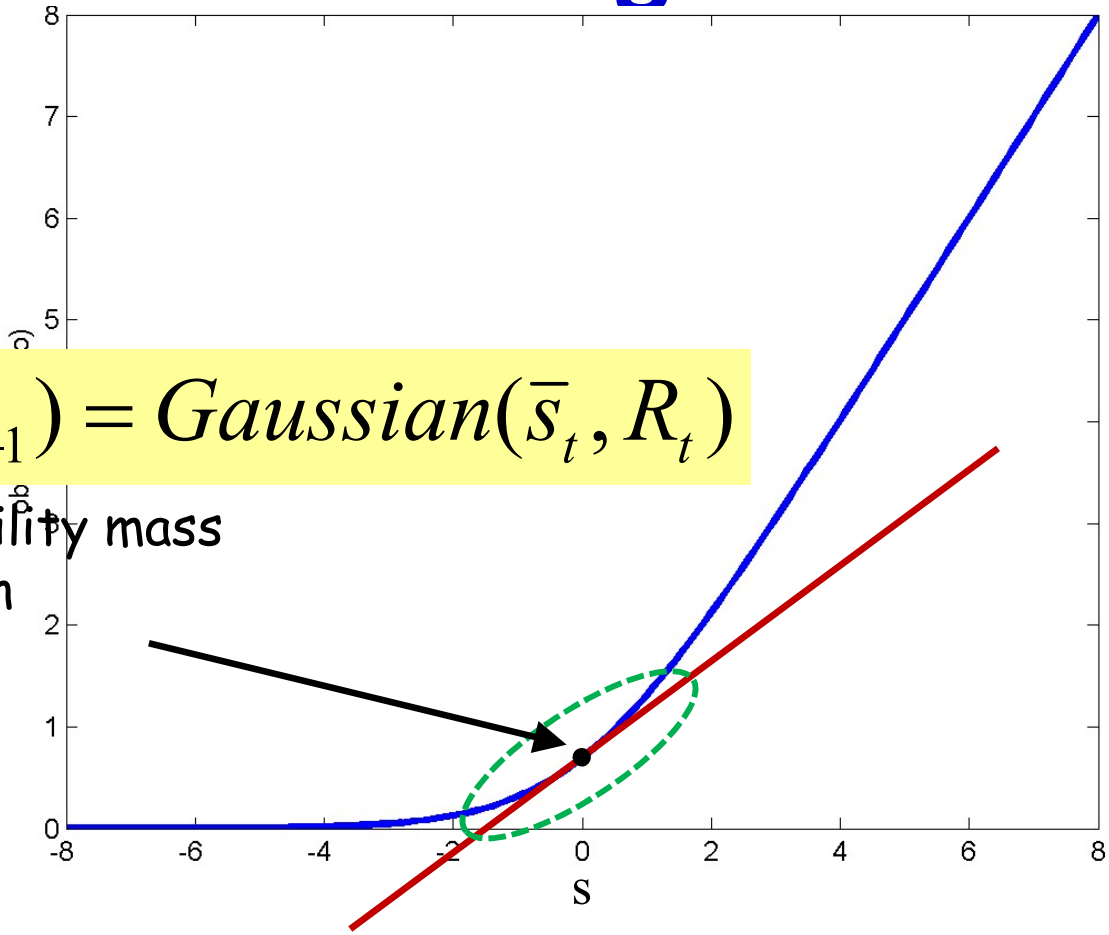
$$o = \gamma + g(s) \quad \Rightarrow \quad o \approx \gamma + g(\bar{s}_t) + J_g(\bar{s}_t)(s - \bar{s}_t)$$

- Simple first-order Taylor series expansion
  - J() is the Jacobian matrix
    - Simply a determinant for scalar state
- Expansion around *current* predicted *a priori* (or predicted) mean of the state
  - Linear approximation changes with time

Most probability is in the low-error region

$$P(s_t \mid o_{0:t-1}) = Gaussian(\bar{s}_t, R_t)$$

Most probability mass close to mean



- $P(s_t)$ is small where approximation error is large
  - Most of the probability mass of *s* is in low-error regions

Linearizing the observation function

$$P(s_t \mid o_{0:t-1}) = Gaussian(\bar{s}_t, R_t)$$

$$o = \gamma + g(s) \quad \Rightarrow \quad o \approx \gamma + g(\bar{s}_t) + J_g(\bar{s}_t)(s - \bar{s}_t)$$

- With the linearized approximation the system becomes "linear"
- The observation PDF becomes Gaussian

$$P(\gamma) = Gaussian(\gamma; 0, \Theta_\gamma)$$

$$P(o \mid s) = Gaussian(o; g(\bar{s}) + J_g(\bar{s})(s - \bar{s}), \Theta_\gamma)$$

The state equation?

$$s_t = f(s_{t-1}) + \varepsilon \qquad P(\varepsilon) = Gaussian(\varepsilon; 0, \Theta_\varepsilon)$$

- Again, direct use of f() can be disastrous
- Solution: Linearize

$$P(s_{t-1} \mid o_{0:t-1}) = Gaussian(s_{t-1}; \hat{s}_{t-1}, \hat{R}_{t-1})$$

$$s_t = f(s_{t-1}) + \varepsilon \quad \Rightarrow \quad s_t \approx \varepsilon + f(\hat{s}_{t-1}) + J_f(\hat{s}_{t-1})(s_{t-1} - \hat{s}_{t-1})$$

- Linearize around the mean of the updated distribution of s at $t-1$
  - Converts the system to a linear one

Linearized System

$$o = \gamma + g(s)$$

$$s_t = f(s_{t-1}) + \varepsilon$$

$$o \approx \gamma + g(\bar{s}_t) + J_g(\bar{s}_t)(s - \bar{s}_t)$$

$$s_t \approx \varepsilon + f(\hat{s}_{t-1}) + J_f(\hat{s}_{t-1})(s_{t-1} - \hat{s}_{t-1})$$

- Now we have a simple time-varying linear system
- Kalman filter equations directly apply

The Extended Kalman filter

$$s_t = f(s_{t-1}) + \varepsilon$$

$$o_t = g(s_t) + \gamma$$

- Prediction

$$\bar{s}_t = f(\hat{s}_{t-1})$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

$$A_t = J_f(\hat{s}_{t-1})$$

$$B_t = J_g(\bar{s}_t)$$

Jacobians used in Linearization

- Update

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma \right)^{-1}$$

$$\hat{s}_t = \bar{s}_t + K_t \left(o_t - g(\bar{s}_t) \right)$$

Assuming ε and γ are 0 mean for simplicity

$$\hat{R}_t = \left(I - K_t B_t \right) R_t$$

The Extended Kalman filter

- Prediction

$$s_t = f(s_{t-1}) + \varepsilon$$

$$\bar{s}_t = f(\hat{s}_{t-1})$$

$$o_t = g(s_t) + \gamma$$

The *predicted* state at time t is obtained simply by propagating the estimated state at t-1 through the state dynamics equation

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma\right)$$

$$\hat{s}_t = \bar{s}_t + K_t\left(o_t - g(\bar{s}_t)\right)$$

$$\hat{R}_t = \left(I - K_t B_t\right) R_t$$

The Extended Kalman filter

- Prediction

$$s_t = f(s_{t-1}) + \varepsilon$$

$$o_t = g(s_t) + \varepsilon$$

$$\bar{s}_t = f(\hat{s}_{t-1})$$

$$A_t = J_f(\hat{s}_{t-1})$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

$$B_t = J_g(\bar{s}_t)$$

Uncertainty of prediction.
The variance of the predictor = variance of ε_t + variance of As_{t-1}

A is obtained by linearizing f()

$$K_t = (I - K_t B_t) R_t$$

The Extended Kalman filter

$$s_t = f(s_{t-1}) + \varepsilon$$

- Prediction

$$\bar{s}_t = f(\hat{s}_{t-1})$$

$$o_t = g(s_t) + \varepsilon$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

$$B_t = J_g(\bar{s}_t)$$

- Update

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma \right)^{-1}$$

The Kalman gain is the slope of the MAP estimator that predicts s from o
RBT = C_{so}, $(BRB^T+\Theta) = C_{oo}$
B is obtained by linearizing g()

The Extended Kalman filter

- Prediction

$$s_t = f(s_{t-1}) + \varepsilon$$

$$\bar{s}_t = f(\hat{s}_{t-1}) \quad \rightarrow \quad o_t = g(s_t) + \varepsilon$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

We can also predict the *observation* from the predicted state using the observation equation

$$\hat{s}_t = \bar{s}_t + K_t\left(o_t - g(\bar{s}_t)\right)$$

$$\hat{R}_t = \left(I - K_t B_t\right) R_t$$

$$\bar{o}_t = g(\bar{s}_t)$$

The Extended Kalman filter

- Prediction

$$s_t = f(s_{t-1}) + \varepsilon$$

$$\bar{s}_t = f(\hat{s}_{t-1})$$

$$o_t = g(s_t) + \varepsilon$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

We must correct the predicted value of the state after making an observation

$$\hat{s}_t = \bar{s}_t + K_t\left(o_t - g(\bar{s}_t)\right)$$

$$\bar{o}_t = g(\bar{s}_t)$$

The correction is the difference between the *actual* observation and the *predicted* observation, scaled by the Kalman Gain

The Extended Kalman filter

- Prediction

$$s_t = f(s_{t-1}) + \varepsilon$$

$$\bar{s}_t = f(\hat{s}_{t-1})$$

$$o_t = g(s_t) + \varepsilon$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

$$B_t = J_g(\bar{s}_t)$$

The uncertainty in state decreases if we observe the data and make a correction

The reduction is a multiplicative "shrinkage" based on Kalman gain and B

$$\hat{R}_t = (I - K_t B_t) R_t$$

The Extended Kalman filter

$$s_t = f(s_{t-1}) + \varepsilon$$

$$o_t = g(s_t) + \varepsilon$$

- Prediction

$$\bar{s}_t = f(\hat{s}_{t-1})$$

$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

$$A_t = J_f(\hat{s}_{t-1})$$

$$B_t = J_g(\bar{s}_t)$$

- Update

$$K_t = R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma \right)^{-1}$$

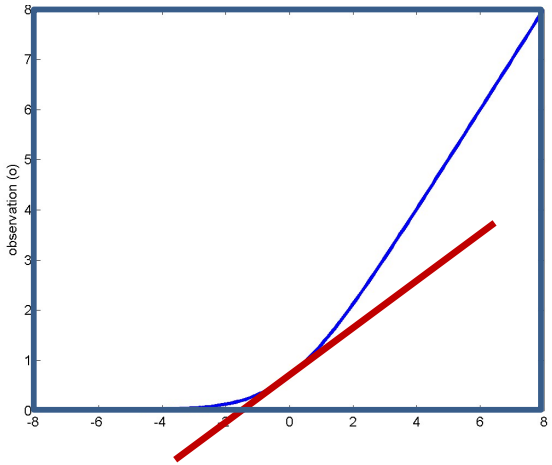
$$\hat{s}_t = \bar{s}_t + K_t \left(o_t - g(\bar{s}_t) \right)$$

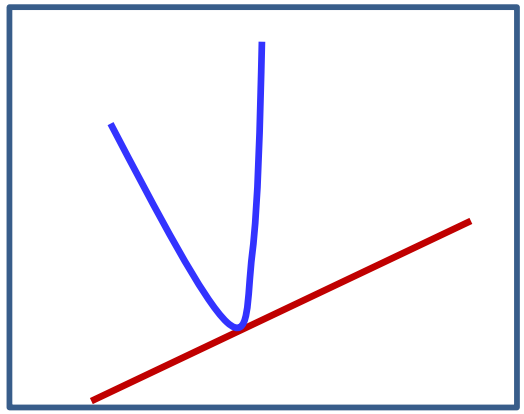
$$\hat{R}_t = \left(I - K_t B_t \right) R_t$$

EKFs

- EKFs are probably the most commonly used algorithm for tracking and prediction
  - Most systems are non-linear
  - Specifically, the relationship between state and observation is usually nonlinear
  - The approach can be extended to include non-linear functions of noise as well
- The term "Kalman filter" often simply refers to an *extended* Kalman filter in most contexts.
- But..

EKFs have limitations





- If the non-linearity changes too quickly with s, the linear approximation is invalid
  - Unstable
- The estimate is often biased
  - The true function lies entirely on one side of the approximation
- Various extensions have been proposed:
  - Invariant extended Kalman filters (IEKF)
  - Unscented Kalman filters (UKF)

Conclusions

- HMMs are predictive models
- Continuous-state models are simple extensions of HMMs
  - Same math applies
- Prediction of linear, Gaussian systems can be performed by Kalman filtering
- Prediction of non-linear, Gaussian systems can be performed by Extended Kalman filtering

A different problem: Non-Gaussian PDFs

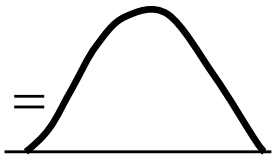
$$o_t = g(s_t) + \gamma \qquad s_t = f(s_{t-1}) + \varepsilon$$

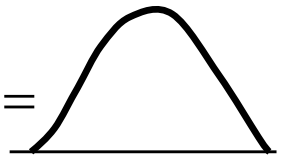
- We have assumed so far that:
  - $P_0(s)$ is Gaussian or can be approximated as Gaussian
  - $P(\varepsilon)$ is Gaussian
  - $P(\gamma)$ is Gaussian
- This has a happy consequence: All distributions remain Gaussian

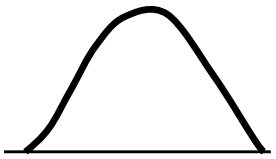
Linear Gaussian Model

$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

$P(s)$ = 
a priori

$P(s_t|s_{t-1})$ = 
Transition prob.

$P(O_t|s_t)$ = 
State output prob

$P(s_0) = P(s)$

$P(s_0|O_0) = C\,P(s_0)\,P(O_0|s_0)$

$$P(s_1|O_0) = \int_{-\infty}^{\infty} P(s_0|O_0)P(s_1|s_0)ds_0$$

$P(s_1|O_{0:1}) = C\,P(s_1|O_0)\,P(O_1|s_0)$

$$P(s_2|O_{0:1}) = \int_{-\infty}^{\infty} P(s_1|O_{0:1})P(s_2|s_1)ds_1$$

$P(s_2|O_{0:2}) = C\,P(s_2|O_{0:1})\,P(O_2|s_2)$

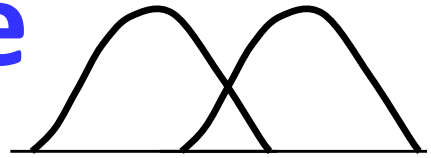
All distributions remain Gaussian

A different problem: Non-Gaussian PDFs

$$o_t = g(s_t) + \gamma \qquad s_t = f(s_{t-1}) + \varepsilon$$

- We have assumed so far that:
  - $P_0(s)$ is Gaussian or can be approximated as Gaussian
  - $P(\varepsilon)$ is Gaussian
  - $P(\gamma)$ is Gaussian
- This has a happy consequence: All distributions remain Gaussian
- But when any of these are not Gaussian, the results are not so happy

A simple case

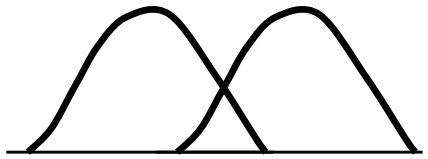


$$o_t = Bs_t + \gamma$$

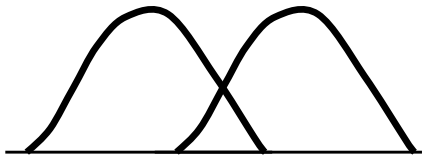
$$P(\gamma) = \sum_{i=0}^{1} w_i Gaussian(\gamma; \mu_i, \Theta_i)$$

- P(γ) is a mixture of only two Gaussians
- *o* is a linear function of *s*
  - Non-linear functions would be linearized anyway
- P(o|s) is also a Gaussian mixture!

$$P(o_t \mid s_t) = P(\gamma = o_t - Bs_t) = \sum_{i=0}^{1} w_i Gaussian(o; \mu_i + Bs_t, \Theta_i)$$

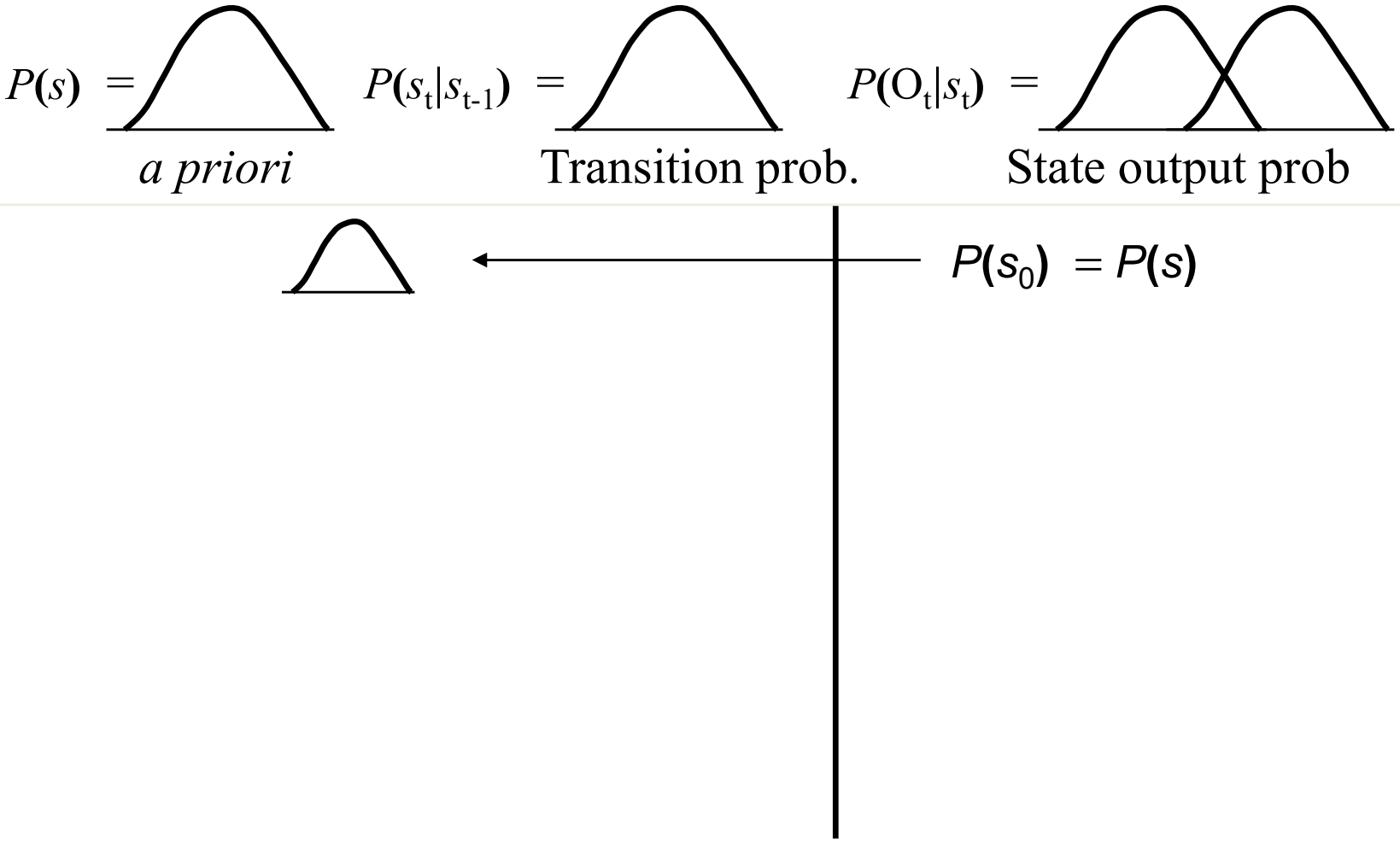


$P(\gamma)$

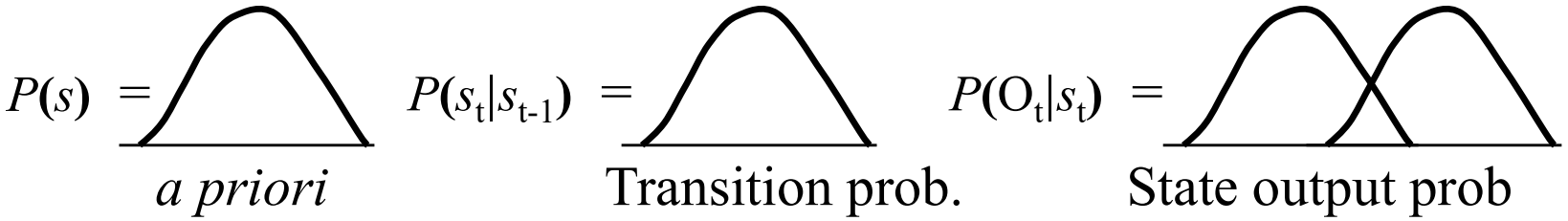


$P(o_t \mid s_t)$

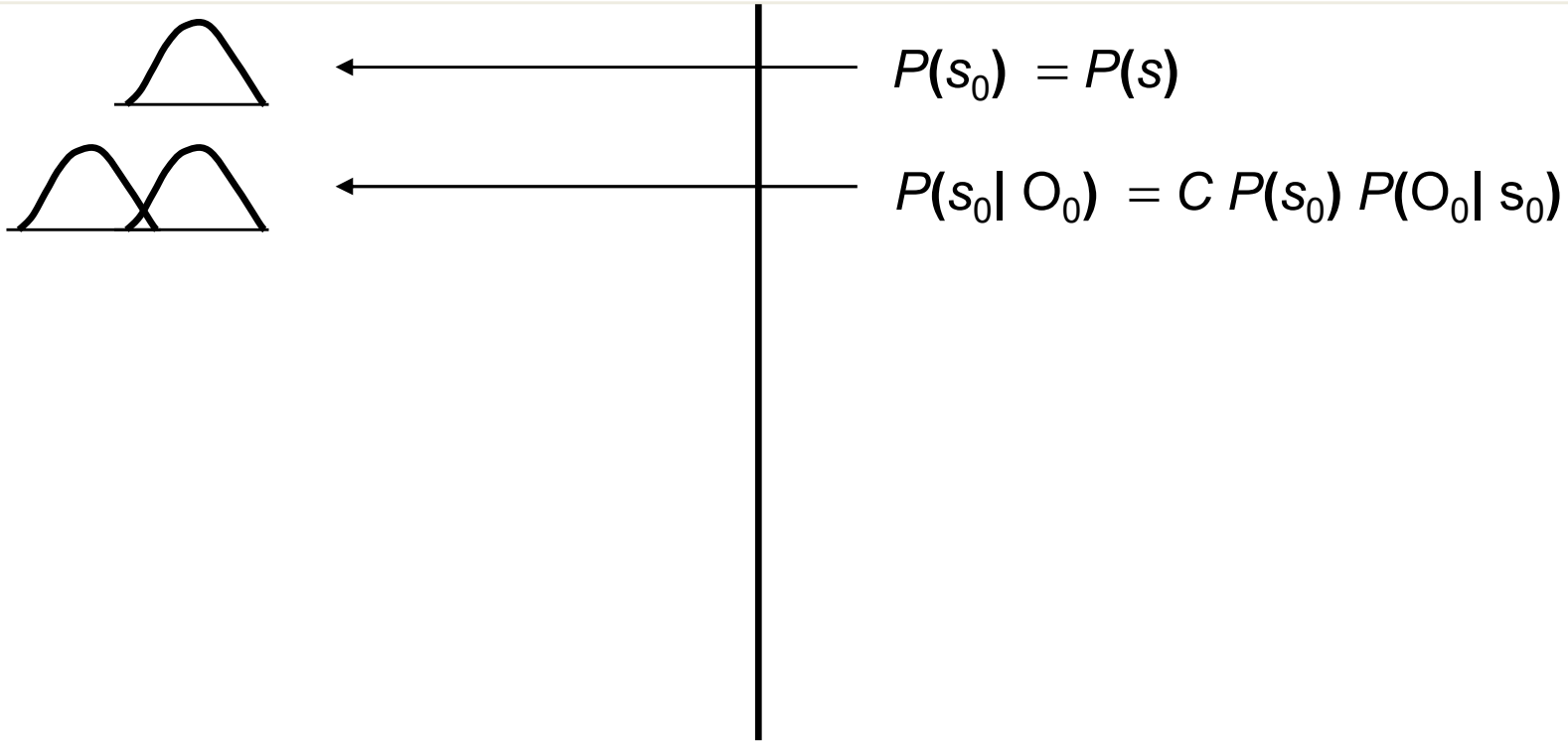
When distributions are not Gaussian



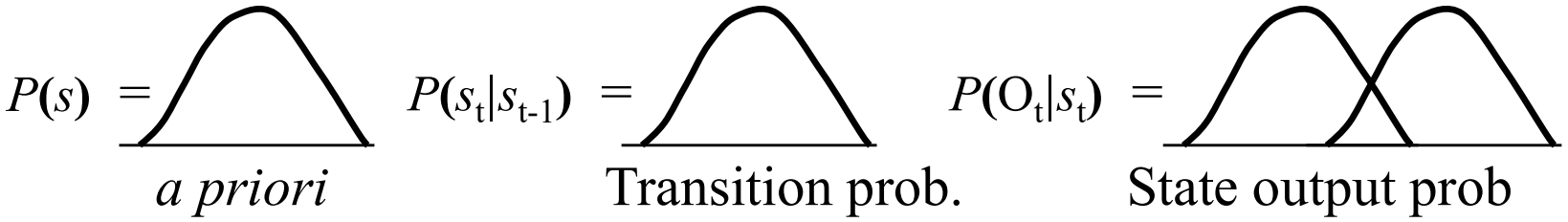
When distributions are not Gaussian



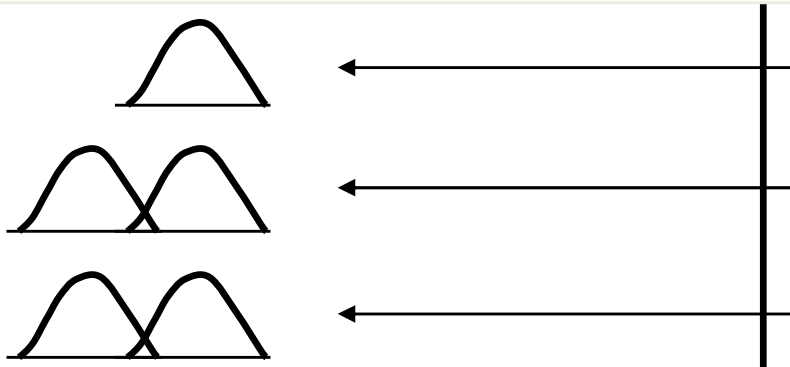
a priori Transition prob. State output prob



When distributions are not Gaussian



a priori Transition prob. State output prob

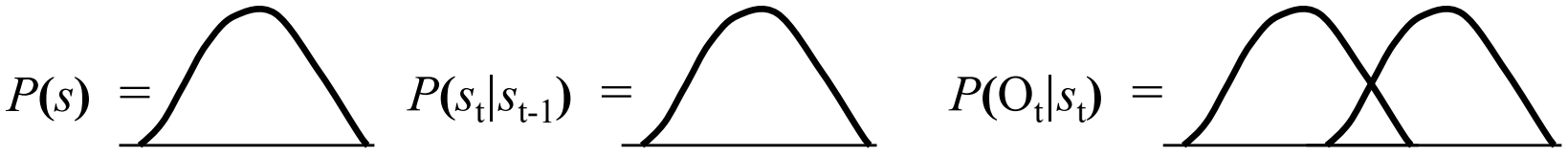


$$P(s_0) = P(s)$$

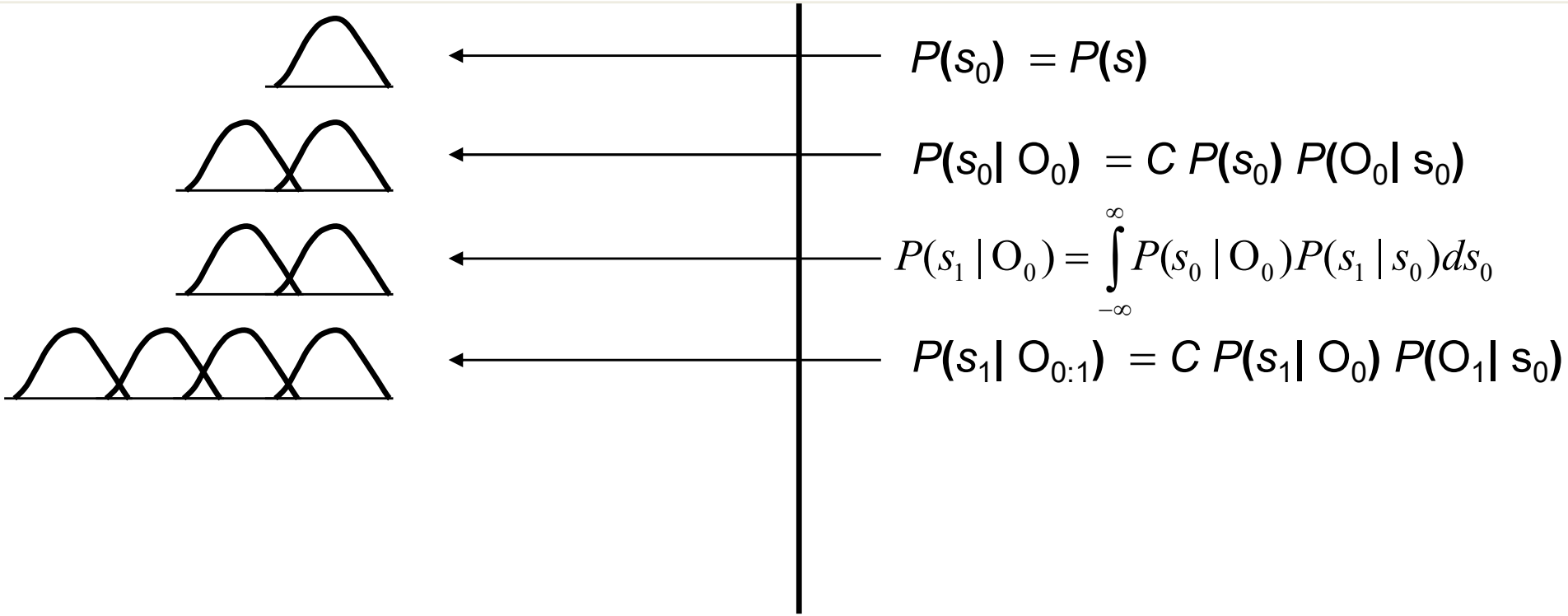
$$P(s_0 | O_0) = C\, P(s_0)\, P(O_0 | s_0)$$

$$P(s_1 | O_0) = \int_{-\infty}^{\infty} P(s_0 | O_0) P(s_1 | s_0) ds_0$$

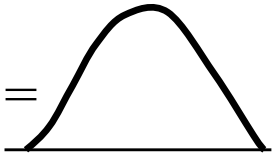
When distributions are not Gaussian



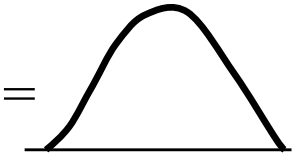
a priori Transition prob. State output prob



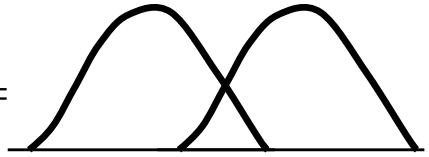
When distributions are not Gaussian

$P(s)$ = 

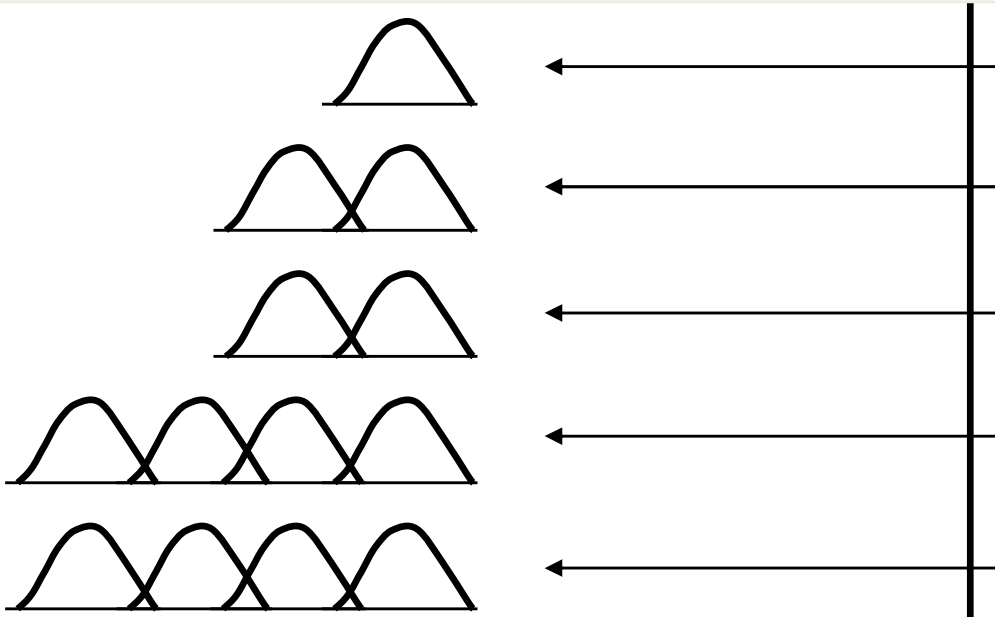
a priori

$P(s_t|s_{t-1})$ = 

Transition prob.

$P(O_t|s_t)$ = 

State output prob



$P(s_0) = P(s)$

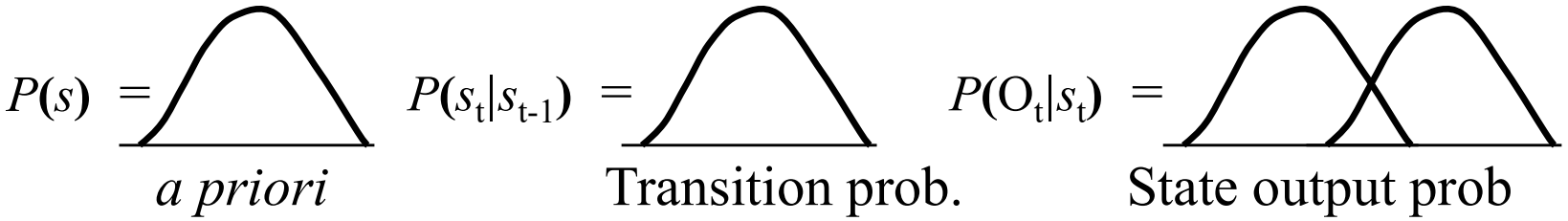
$P(s_0|O_0) = C\,P(s_0)\,P(O_0|s_0)$

$$P(s_1 \mid O_0) = \int_{-\infty}^{\infty} P(s_0 \mid O_0) P(s_1 \mid s_0) ds_0$$

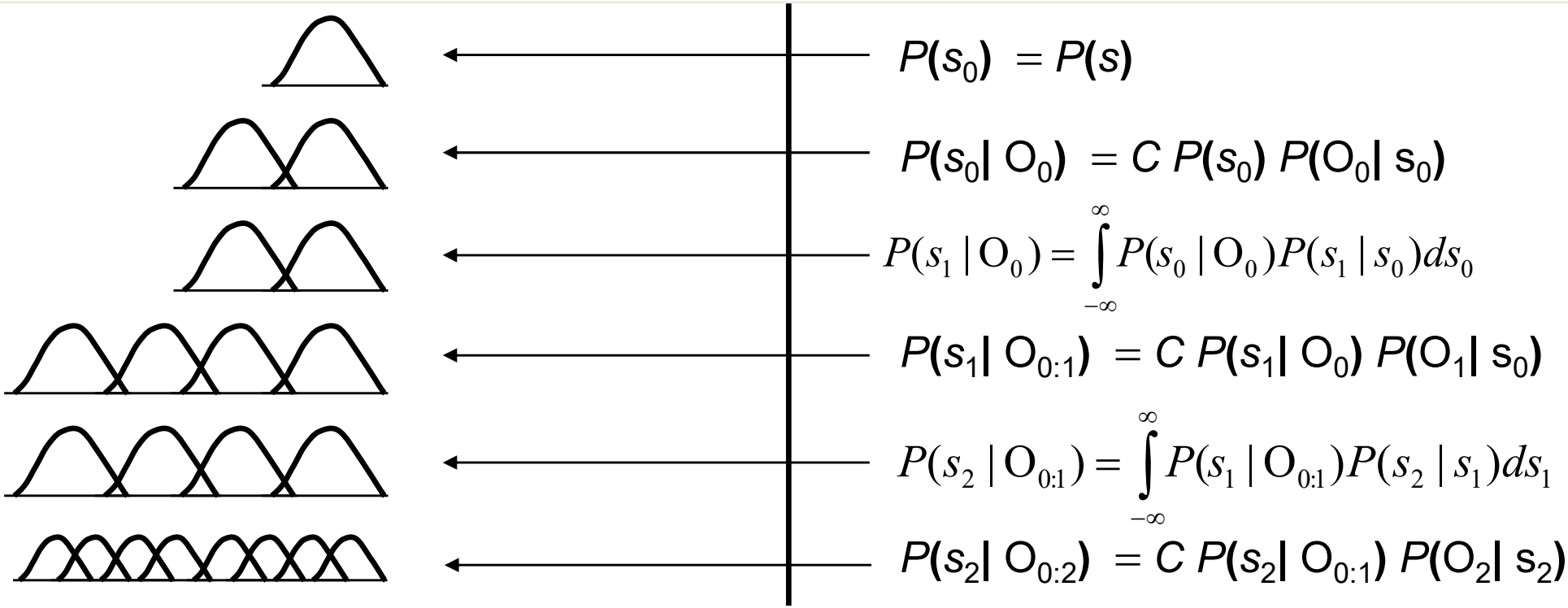
$P(s_1|O_{0:1}) = C\,P(s_1|O_0)\,P(O_1|s_0)$

$$P(s_2 \mid O_{0:1}) = \int_{-\infty}^{\infty} P(s_1 \mid O_{0:1}) P(s_2 \mid s_1) ds_1$$

When distributions are not Gaussian



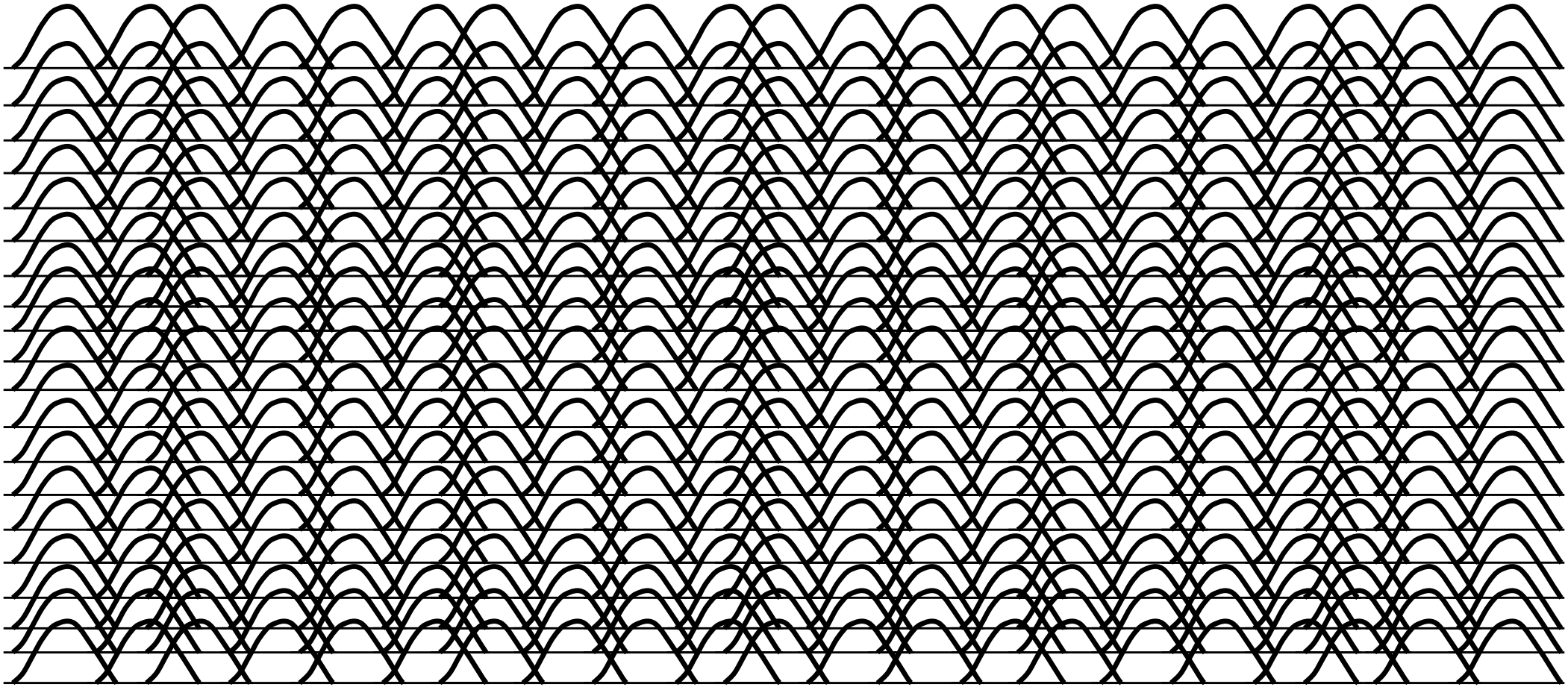
a priori Transition prob. State output prob



When $P(O_t|s_t)$ has more than one Gaussian, after only a few time steps…

When distributions are not Gaussian

$P(s_t \mid \mathrm{O}_{0:t}) =$



We have too many Gaussians for comfort..

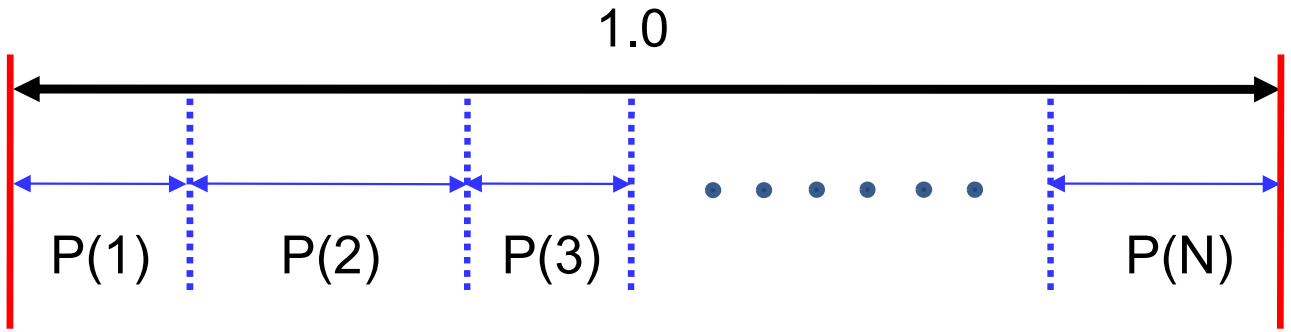
Related Topic: How to sample from a Distribution?

- "Sampling from a Distribution P(x; Γ) with parameters Γ"
- Generate random numbers such that
  - The distribution of a large number of generated numbers is P(x; Γ)
  - The parameters of the distribution are Γ
- Many algorithms to generate RVs from a variety of distributions
  - Generation from a uniform distribution is well studied
  - Uniform RVs used to sample from multinomial distributions
  - Other distributions: Most commonly, transform a uniform RV to the desired distribution

Sampling from a multinomial

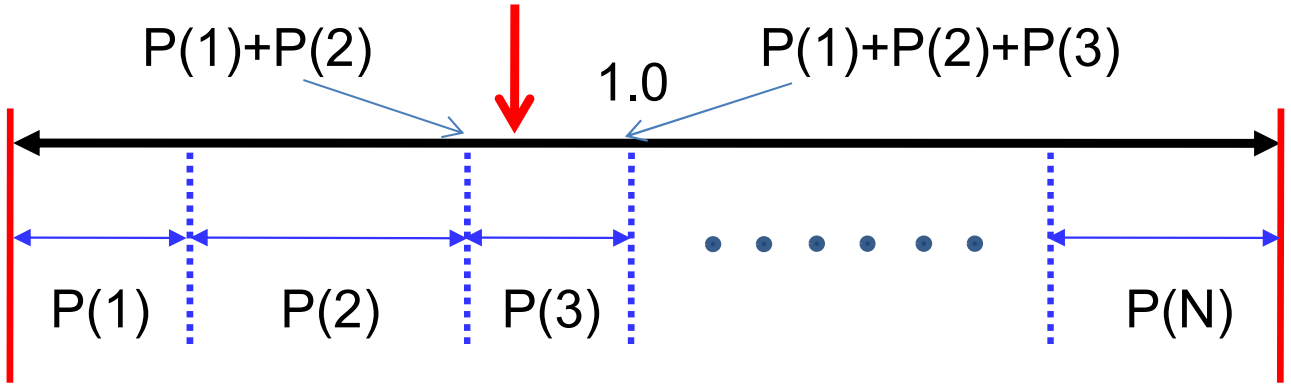
- Given a multinomial over N symbols, with probability of i^{th} symbol = P(i)
- Randomly generate symbols from this distribution
- Can be done by sampling from a uniform distribution

Sampling a multinomial



- Segment a range (0,1) according to the probabilities P(i)
  - The P(i) terms will sum to 1.0

Sampling a multinomial



- Segment a range (0,1) according to the probabilities P(i)
  - The P(i) terms will sum to 1.0
- Randomly generate a number from a uniform distribution
  - Matlab: "rand".
  - Generates a number between 0 and 1 with uniform probability
- If the number falls in the i^{th} segment, select the i^{th} symbol

Related Topic: Sampling from a Gaussian

- Many algorithms
  - Simplest: add many samples from a uniform RV
  - The sum of 12 uniform RVs (uniform in (0,1)) is approximately Gaussian with mean 6 and variance 1
  - For scalar Gaussian, mean μ, std dev σ:

$$x = \sum_{i=1}^{12} r_i - 6$$

- Matlab : x = μ + randn* σ
  - "randn" draws from a Gaussian of mean=0, variance=1

Related Topic: Sampling from a Gaussian

- Multivariate (d-dimensional) Gaussian with mean μ and covariance Θ
  - Compute eigen value matrix Λ and eigenvector matrix E for Θ
    - $\Theta = E \Lambda E^T$
  - Generate d 0-mean unit-variance numbers $x_1..x_d$
  - Arrange them in a vector:

$$X = [x_1 .. x_d]^T$$

  - Multiply X by the square root of Λ and E, add μ

$$Y = \mu + E \, sqrt(\Lambda) \, X$$

Sampling from a Gaussian Mixture

$$\sum_i w_i Gaussian(X; \mu_i, \Theta_i)$$

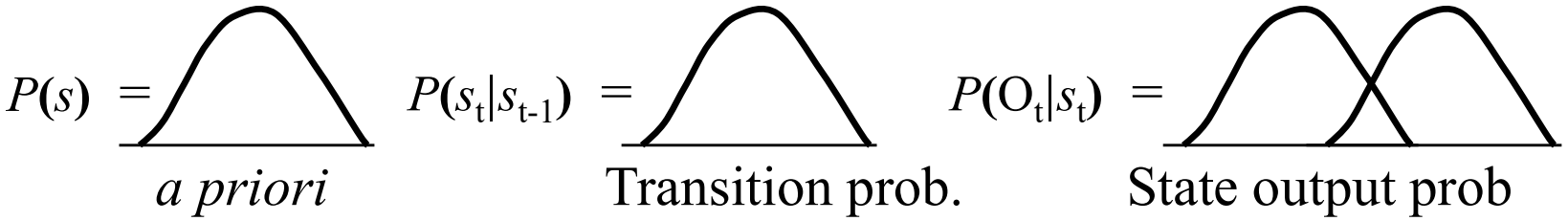
- Select a Gaussian by sampling the multinomial distribution of weights:

$$j \sim \text{Category}(w_1, w_2, \ldots)$$

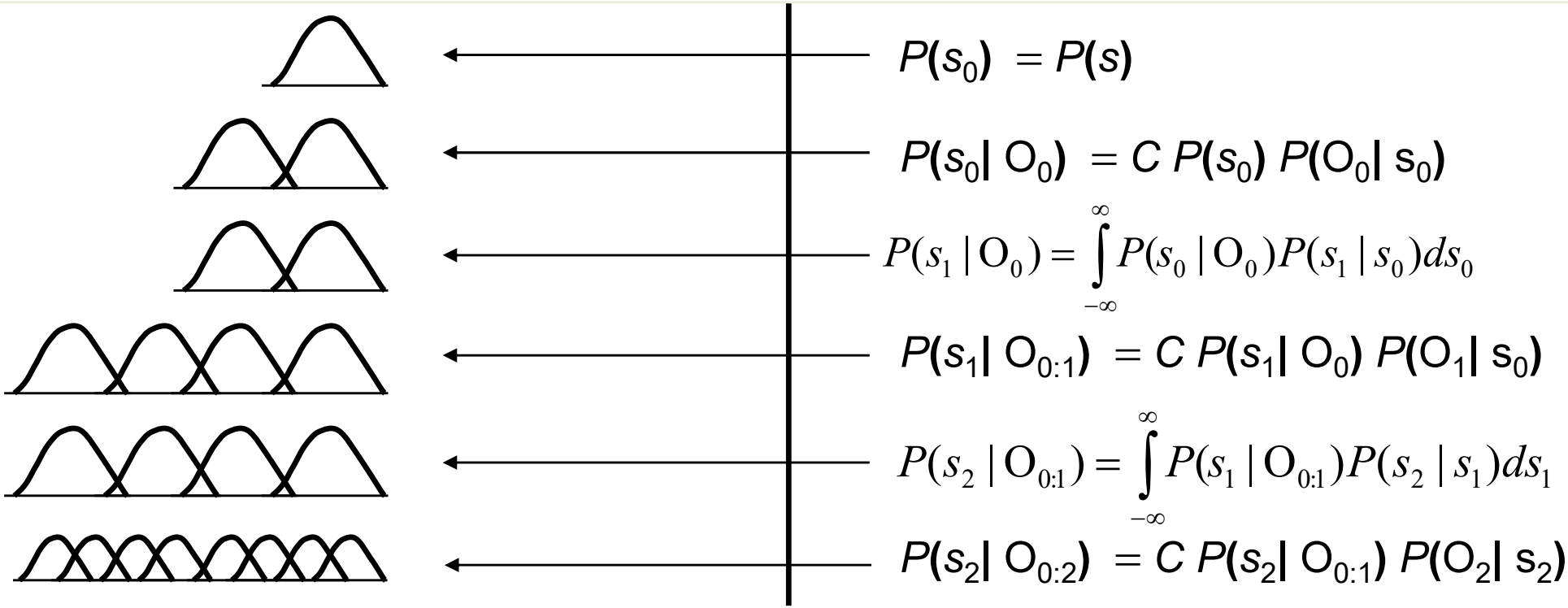
- Sample from the selected Gaussian

$$Gaussian(X; \mu_j, \Theta_j)$$

When distributions are not Gaussian



a priori Transition prob. State output prob

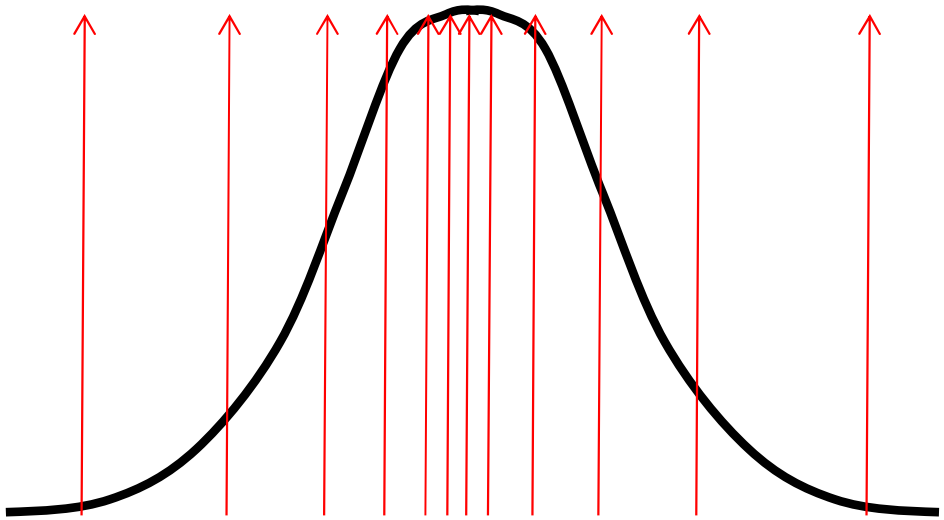


When $P(O_t|s_t)$ has more than one Gaussian, after only a few time steps…

The problem of the exploding distribution

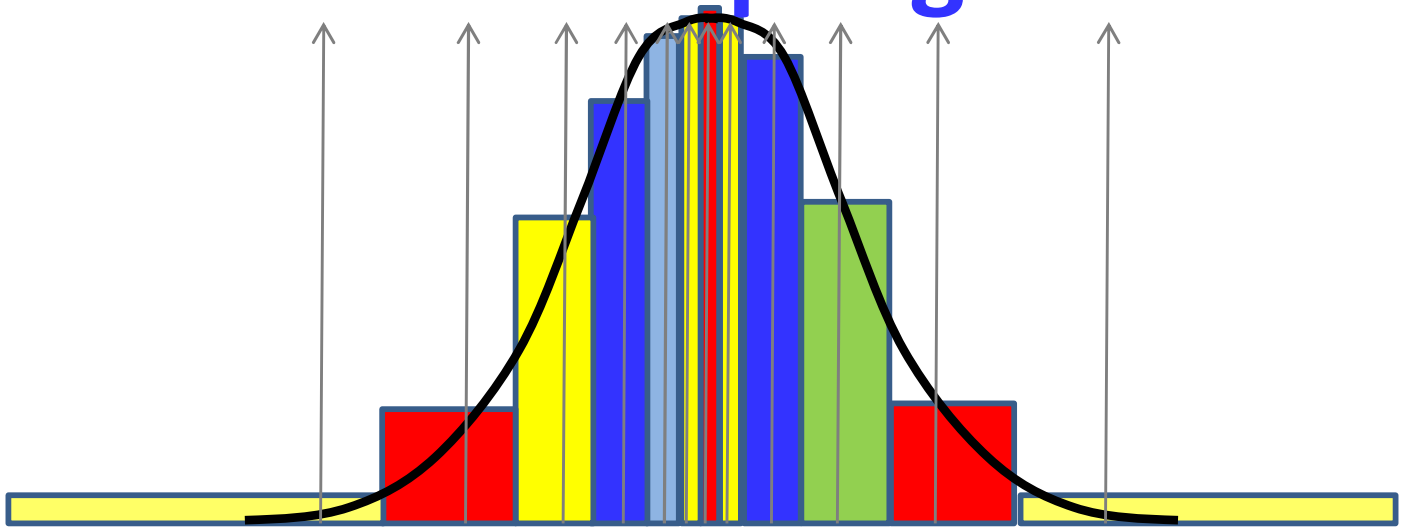
- The complexity of the distribution increases exponentially with time
- This is a consequence of having a *continuous* state space
  - Only Gaussian PDFs propagate without increase of complexity
- *Discrete-state* systems do not have this problem
  - The number of states in an HMM stays fixed
  - However, discrete state spaces are too coarse
- Solution: Combine the two concepts
  - *Discretize* the state space dynamically

Discrete approximation to a distribution



- A large-enough collection of randomly-drawn samples from a distribution will approximately quantize the space of the random variable into equi-probable regions
  - We have more random samples from high-probability regions and fewer samples from low-probability reigons

Discrete approximation: Random sampling



- A PDF can be approximated as a uniform probability distribution over randomly drawn samples
  - Since each sample represents approximately the same probability mass (1/M if there are M samples)

$$P(x) \approx \frac{1}{M} \sum_{i=0}^{M-1} \delta(x - x_i)$$

Note: Properties of a discrete distribution

$$P(x) \approx \frac{1}{M}\sum_{i=0}^{M-1}\delta(x - x_i)$$

$$P(x)P(y \mid x) \propto \sum_{i=0}^{M-1} P(y \mid x_i)\delta(x - x_i)$$

- The product of a discrete distribution with another distribution is simply a weighted discrete probability

$$P(x) \approx \sum_{i=0}^{M-1} w_i\delta(x - x_i)$$

$$\int_{-\infty}^{\infty} P(x)P(y \mid x)dx = \sum_{i=0}^{M-1} w_i P(y \mid x_i)$$

- The integral of the product is a mixture distribution

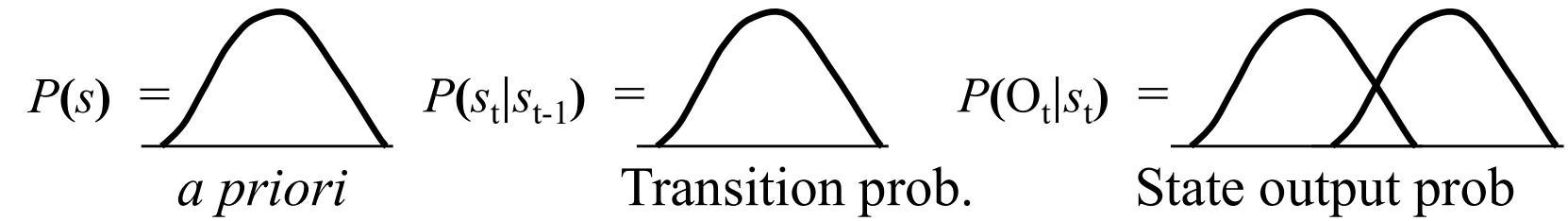
Discretizing the state space

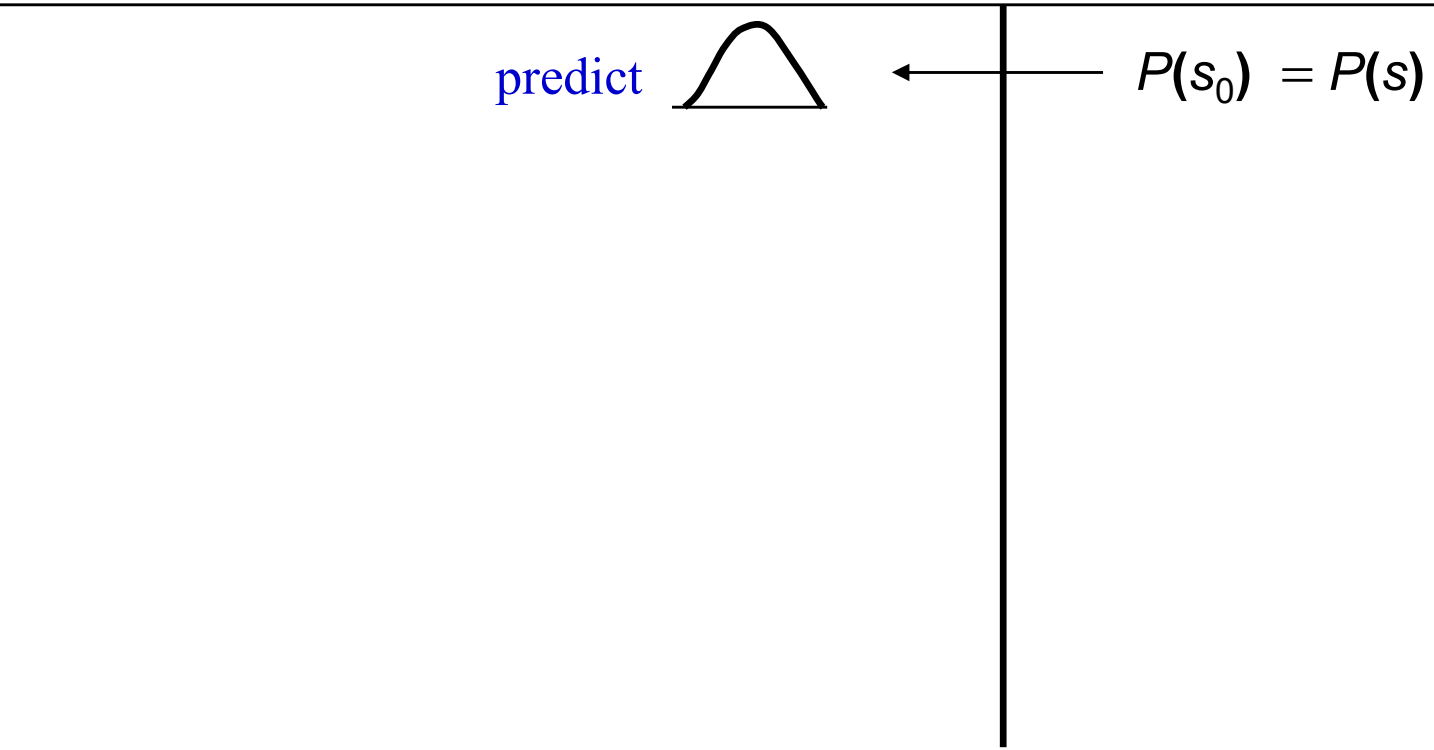
- At each time, discretize the predicted state space

$$P(s_t \mid o_{0:t}) \approx \frac{1}{M} \sum_{i=0}^{M-1} \delta(s_t - s_i)$$

  - s_i are randomly drawn samples from $P(s_t \mid o_{0:t})$
- Propagate the discretized distribution

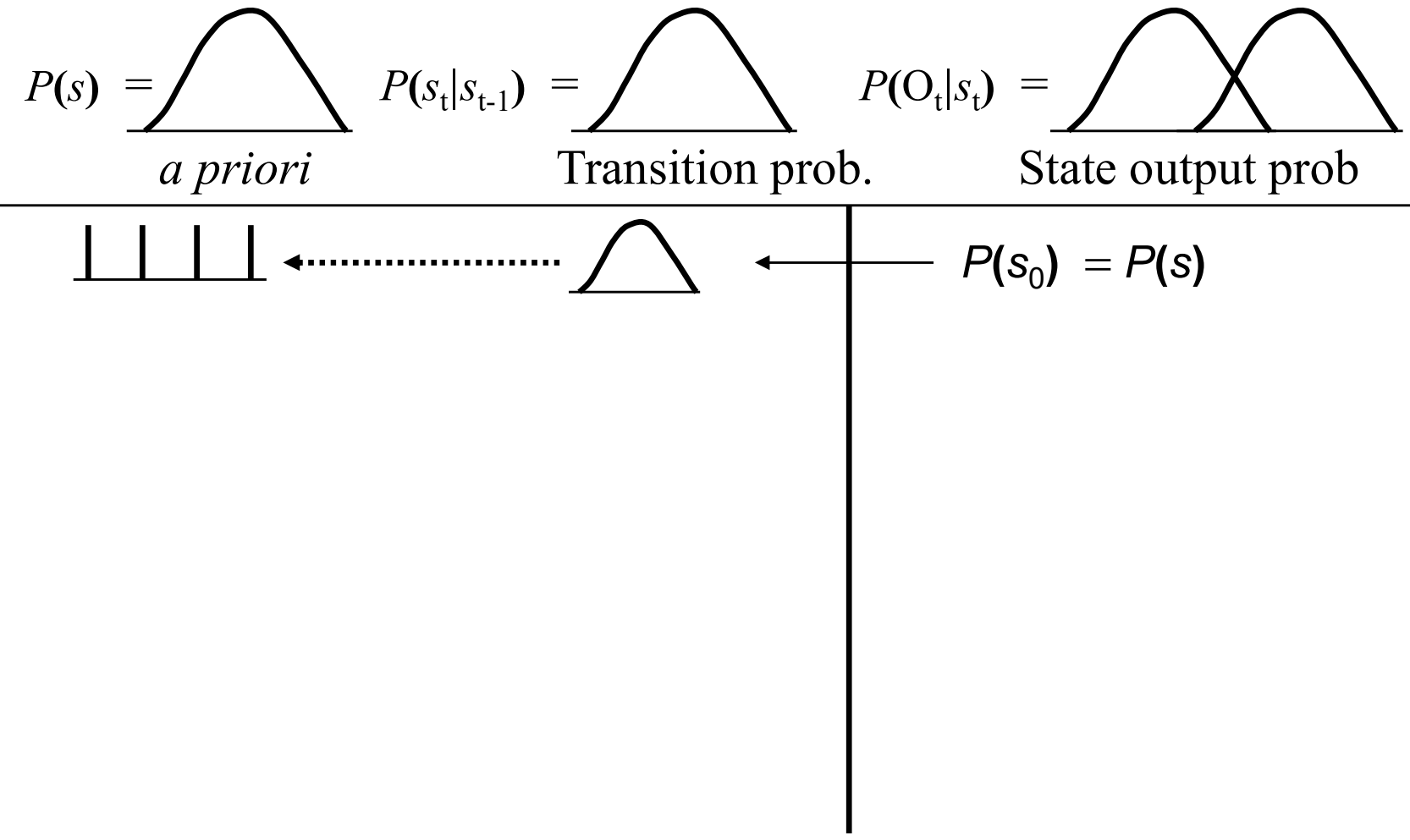
Particle Filtering





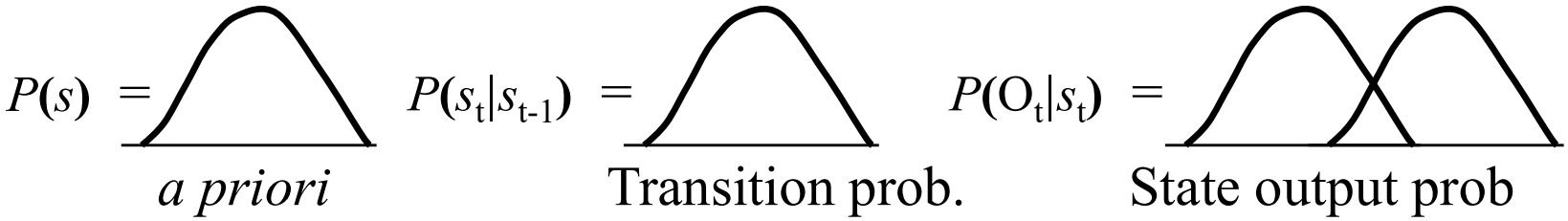
Assuming that we only generate ***FOUR*** samples from the predicted distributions

Particle Filtering

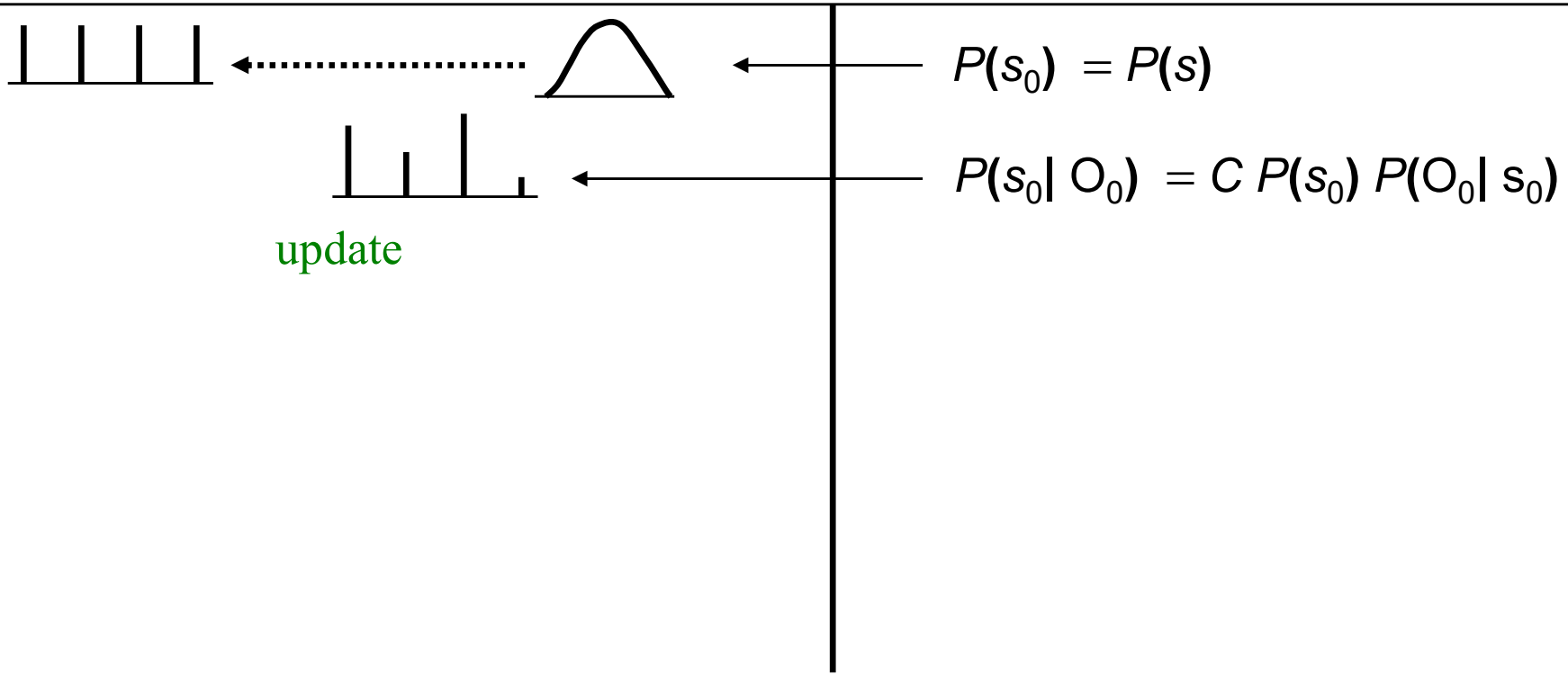


Assuming that we only generate ***FOUR*** samples from the predicted distributions

Particle Filtering

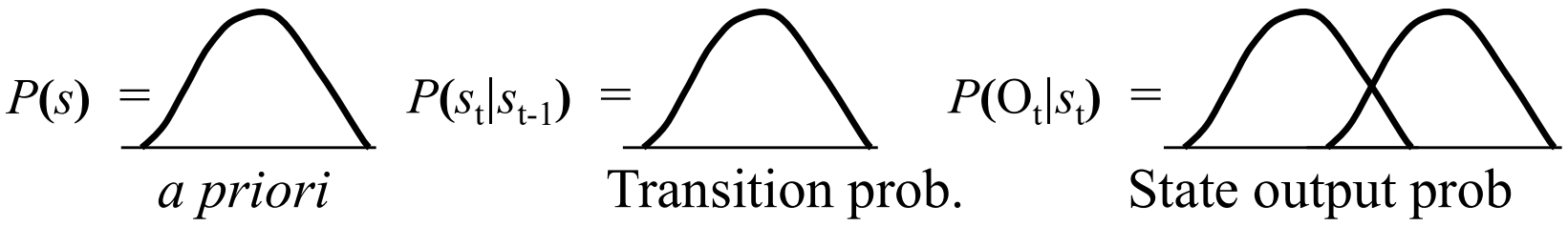


a priori Transition prob. State output prob

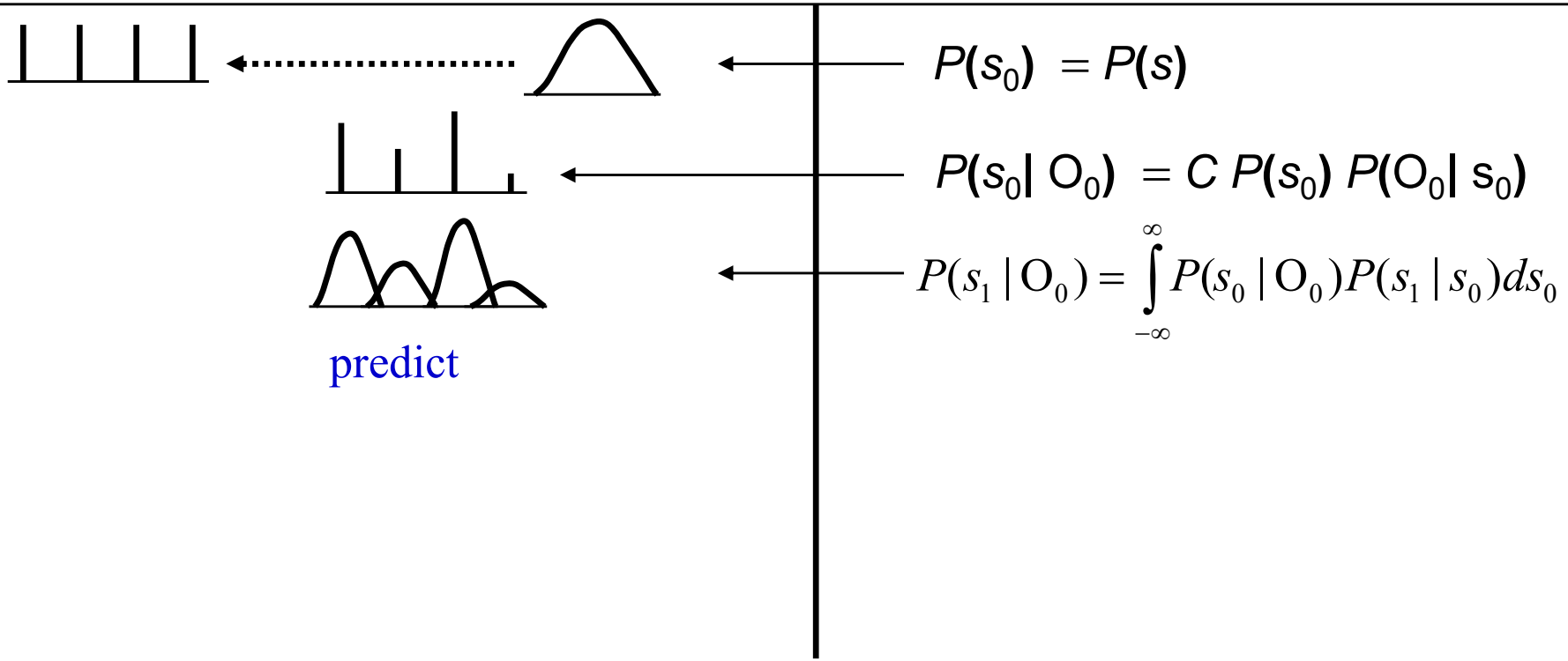


Assuming that we only generate ***FOUR*** samples from the predicted distributions

Particle Filtering

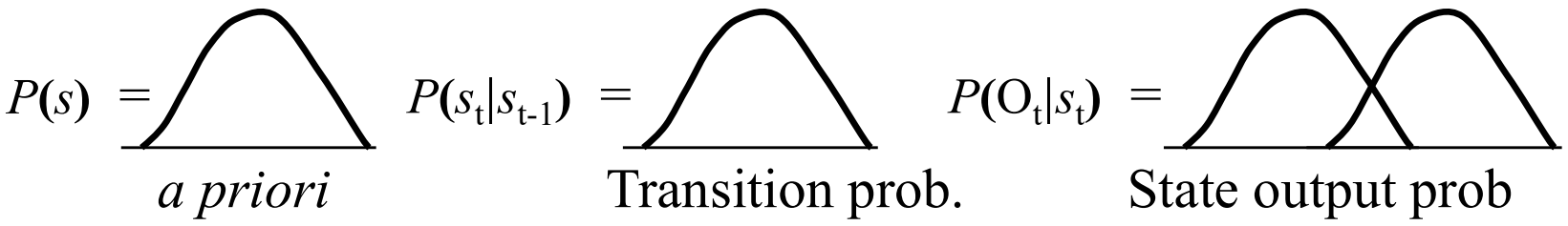


a priori Transition prob. State output prob

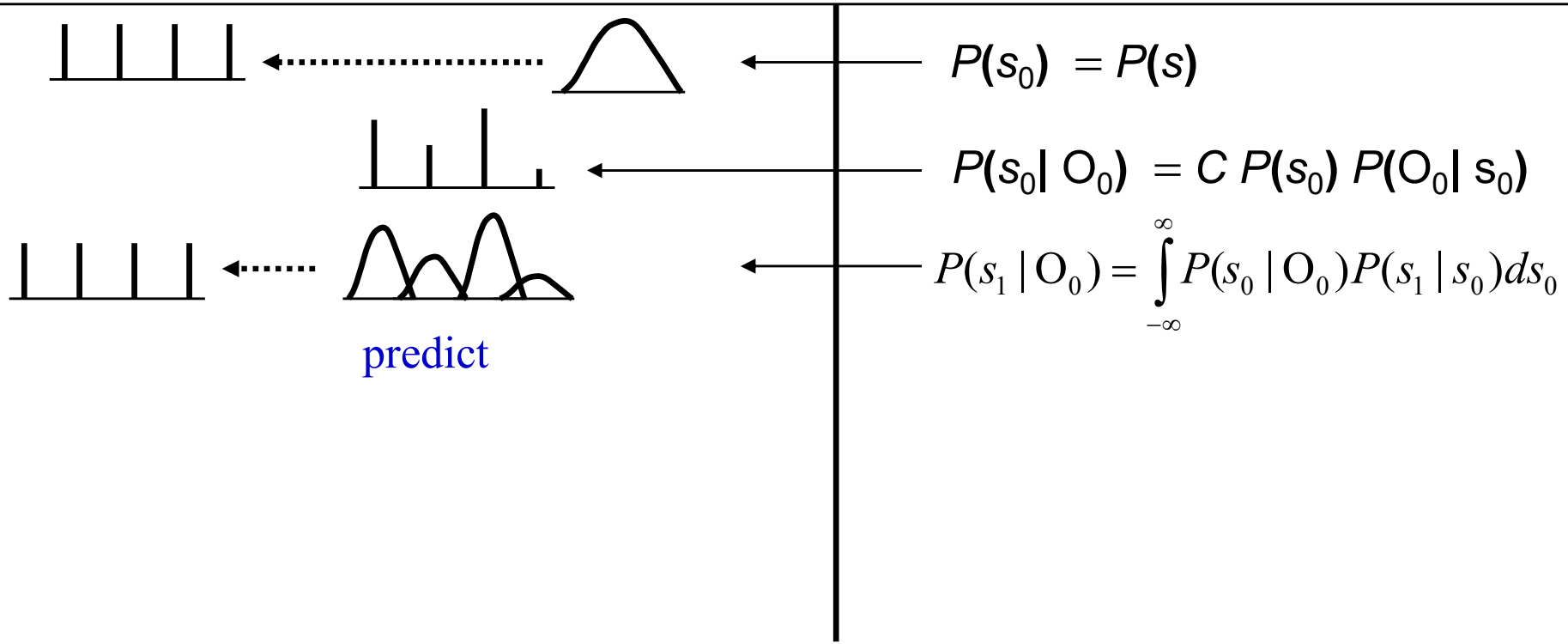


Assuming that we only generate ***FOUR*** samples from the predicted distributions

Particle Filtering

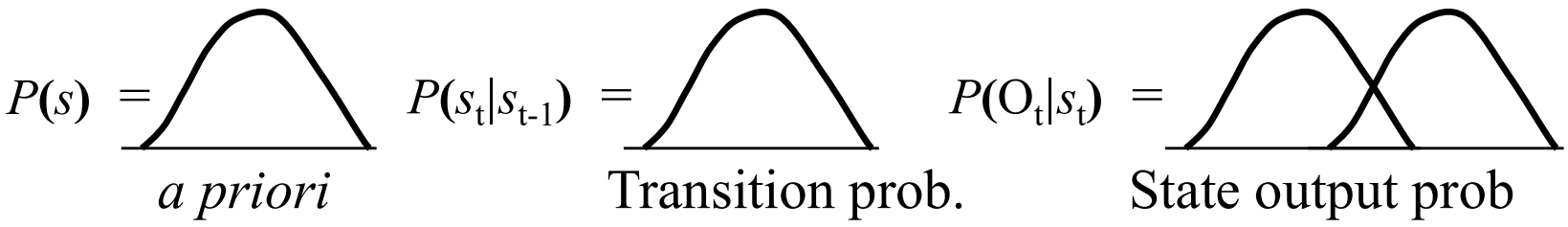


a priori Transition prob. State output prob

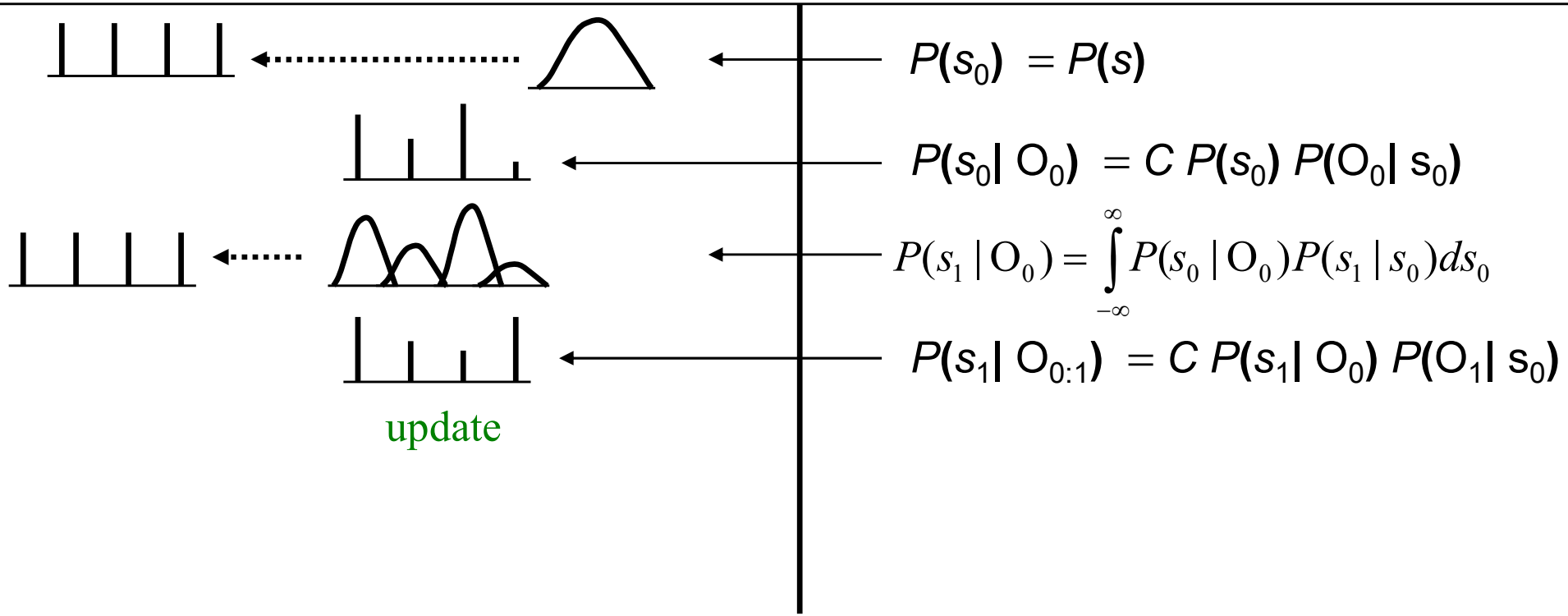


Assuming that we only generate ***FOUR*** samples from the predicted distributions

Particle Filtering

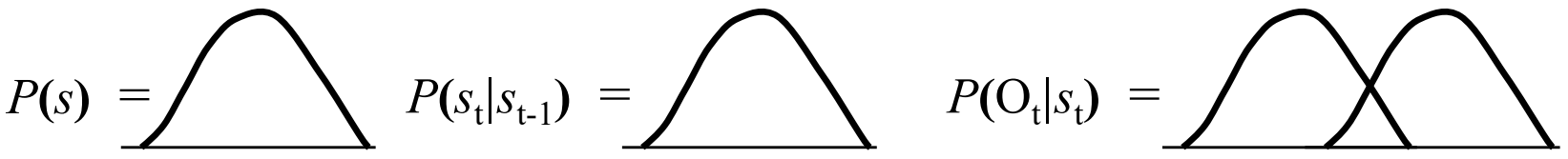


a priori Transition prob. State output prob

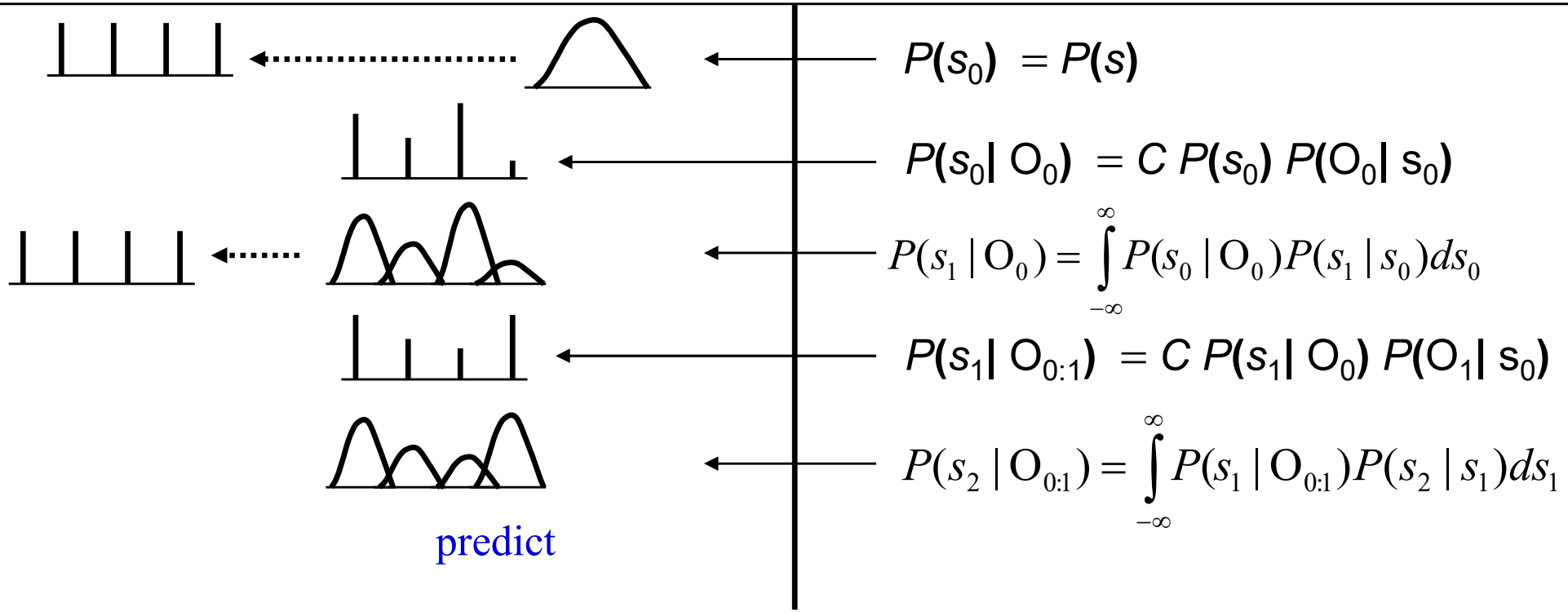


Assuming that we only generate ***FOUR*** samples from the predicted distributions

Particle Filtering

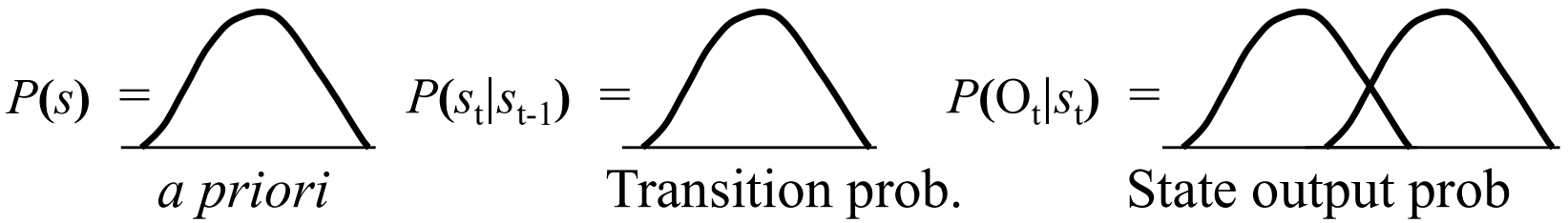


a priori Transition prob. State output prob

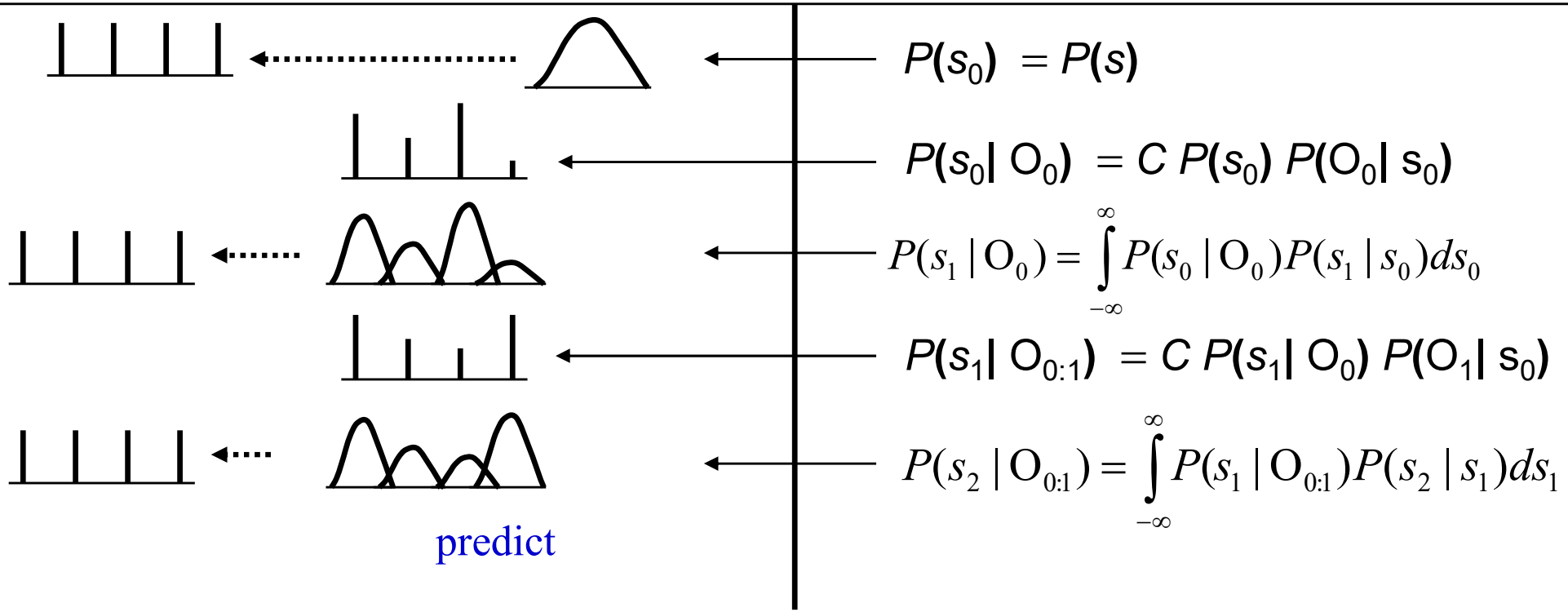


Assuming that we only generate ***FOUR*** samples from the predicted distributions

Particle Filtering

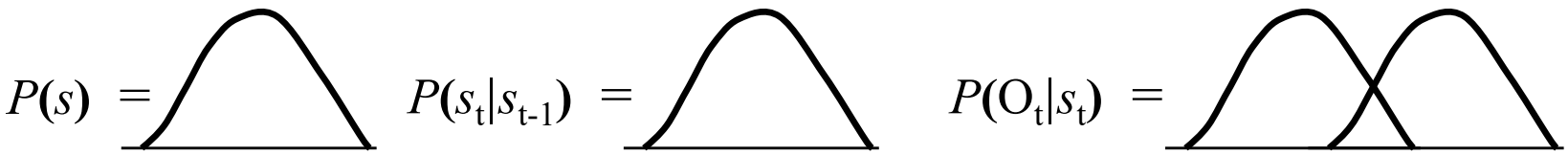


a priori Transition prob. State output prob

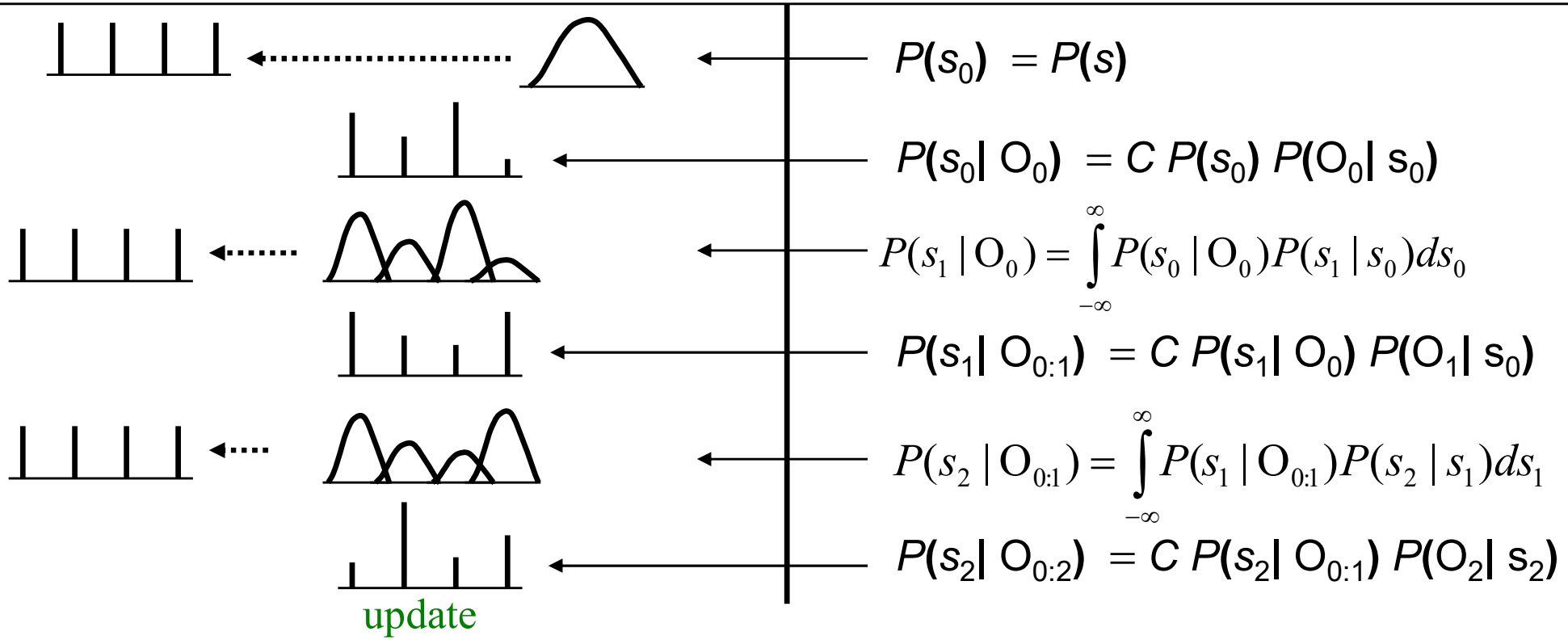


Assuming that we only generate ***FOUR*** samples from the predicted distributions

Particle Filtering



a priori Transition prob. State output prob

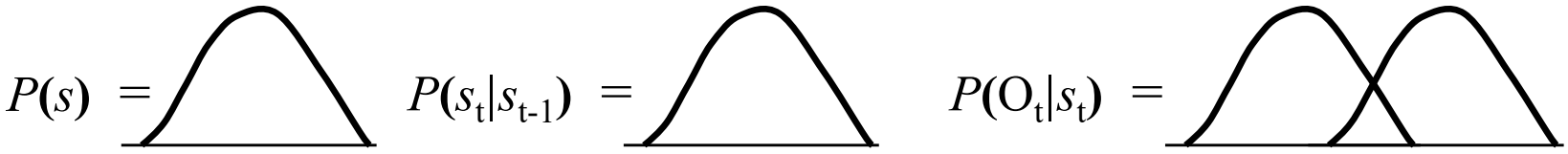


Assuming that we only generate ***FOUR*** samples from the predicted distributions

Particle Filtering

- Discretize state space at the prediction step
  - By sampling the continuous predicted distribution
    - If appropriately sampled, all generated samples may be considered to be equally probable
  - Sampling results in a **discrete** uniform distribution
- Update step updates the distribution of the quantized state space
  - Results in a **discrete** non-uniform distribution
- Predicted state distribution for the next time instant will again be continuous
  - Must be **discretized** again by sampling
- At any step, the current state distribution will not have more components than the number of samples generated at the previous sampling step
  - The complexity of distributions remains constant

Particle Filtering



a priori Transition prob. State output prob

predict

Prediction at time t:

$$P(s_t \mid O_{0:t-1}) = \int_{-\infty}^{\infty} P(s_{t-1} \mid O_{0:t-1}) P(s_t \mid s_{t-1}) ds_{t-1}$$

Update at time t:

update

$$P(s_t \mid O_{0:t}) = CP(s_t \mid O_{0:t-1}) P(O_t \mid s_t)$$

Number of mixture components in predicted distribution governed by number of samples in discrete distribution

By deriving a small (100-1000) number of samples at each time instant, all distributions are kept manageable

Particle Filtering

$$o_t = g(s_t) + \gamma \qquad s_t = f(s_{t-1}) + \varepsilon$$

$$P_\gamma(\gamma) \qquad P_\varepsilon(\varepsilon)$$

- At t = 0, sample the initial state distribution

$$P(s_0 \mid o_{-1}) = P(s_0) \approx \frac{1}{M}\sum_{i=0}^{M-1}\delta(s_0 - \bar{s}_i^0) \text{ where } \bar{s}_i^0 \leftarrow P_0(s)$$

- Update the state distribution with the observation

$$P(s_t \mid o_{0:t}) = C\sum_{i=0}^{M-1} P_\gamma(o_t - g(\bar{s}_i^t))\delta(s_t - \bar{s}_i^t)$$

$$C = \frac{1}{\sum_{i=0}^{M-1} P_\gamma(o_t - g(\bar{s}_i^t))}$$

Particle Filtering

$$o_t = g(s_t) + \gamma$$

$$s_t = f(s_{t-1}) + \varepsilon$$

$$P_\gamma(\gamma)$$

$$P_\varepsilon(\varepsilon)$$

- Predict the state distribution at the next time

$$P(s_t \mid o_{0:t-1}) = C \sum_{i=0}^{M-1} P_\gamma(o_{t-1} - g(\bar{s}_i^{t-1})) P_\varepsilon(s_t - f(\bar{s}_i^{t-1}))$$

- Sample the predicted state distribution

$$P(s_t \mid o_{0:t-1}) \approx \frac{1}{M} \sum_{i=0}^{M-1} \delta(s_t - \bar{s}_i^t) \text{ where } \bar{s}_i^t \leftarrow P(s_t \mid o_{0:t-1})$$

Particle Filtering

$$o_t = g(s_t) + \gamma \quad P_\gamma(\gamma) \qquad s_t = f(s_{t-1}) + \varepsilon \quad P_\varepsilon(\varepsilon)$$

- Predict the state distribution at t

$$P(s_t \mid o_{0:t-1}) = C \sum_{i=0}^{M-1} P_\gamma(o_{t-1} - g(\bar{s}_i^{t-1})) P_\varepsilon(s_t - f(\bar{s}_i^{t-1}))$$

- Sample the predicted state distribution at t

$$P(s_t \mid o_{0:t-1}) \approx \frac{1}{M} \sum_{i=0}^{M-1} \delta(s_t - \bar{s}_i^t) \text{ where } \bar{s}_i^t \leftarrow P(s_t \mid o_{0:t-1})$$

- Update the state distribution at t

$$P(s_t \mid o_{0:t}) = C \sum_{i=0}^{M-1} P_\gamma(o_t - g(\bar{s}_i^t)) \delta(s_t - \bar{s}_i^t)$$

$$C = \frac{1}{\sum_{i=0}^{M-1} P_\gamma(o_t - g(\bar{s}_i^t))}$$

Estimating a state

- The algorithm gives us a discrete updated distribution over states:

$$P(s_t \mid o_{0:t}) = C \sum_{i=0}^{M-1} P_\gamma(o_t - g(\bar{s}_i^t))\delta(s_t - \bar{s}_i^t)$$

- The actual state can be estimated as the mean of this distribution

$$\hat{s}_t = C \sum_{i=0}^{M-1} \bar{s}_i^t P_\gamma(o_t - g(\bar{s}_i^t))$$

- Alternately, it can be the most likely sample

$$\hat{s}_t = \bar{s}_j^t \;:\; j = \arg\max_i P_\gamma(o_t - g(\bar{s}_i^t))$$

Simulations with a Linear Model

$$s_t = s_{t-1} + \varepsilon_t \qquad o_t = s_t + x_t$$

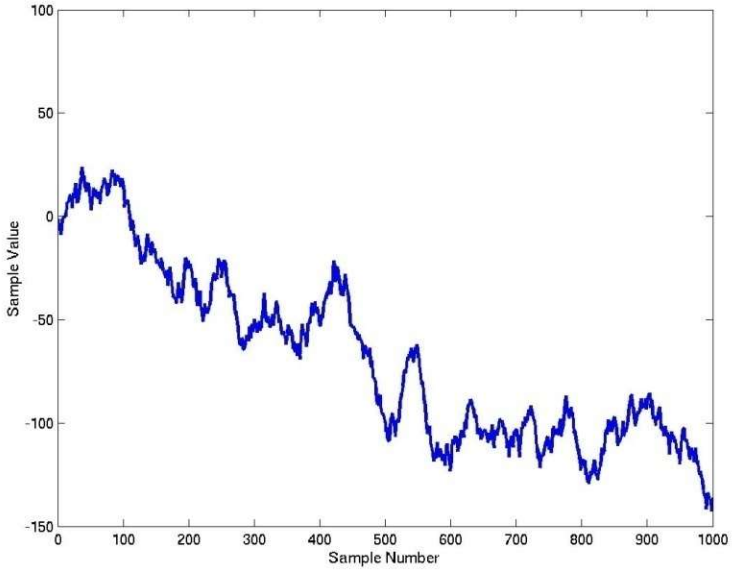
- ε_t has a Gaussian distribution with 0 mean, known variance
- x_t has a mixture Gaussian distribution with known parameters
- Simulation:
  - Generate state sequence S_t from model
  - Generate sequence of X_t from model with one X_t term for every S_t term
  - Generate observation sequence O_t from S_t and X_t
  - Attempt to estimate S_t from O_t

Simulation: Synthesizing data

Generate state sequence according to:

$$s_t = s_{t-1} + \varepsilon_t$$

ε_t is Gaussian with mean 0 and variance 10



Simulation: Synthesizing data

Generate state sequence according to: $s_t = s_{t-1} + \varepsilon_t$

ε_t is Gaussian with mean 0 and variance 10

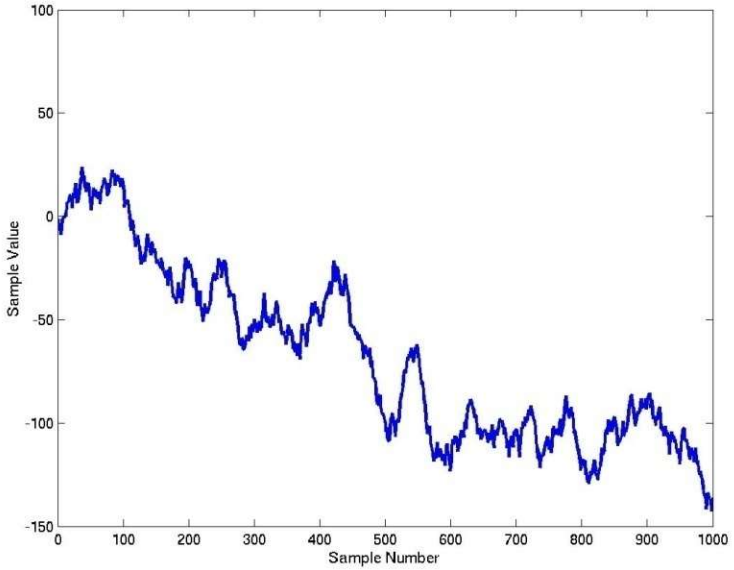
Generate observation sequence from state sequence according to: $o_t = s_t + x_t$

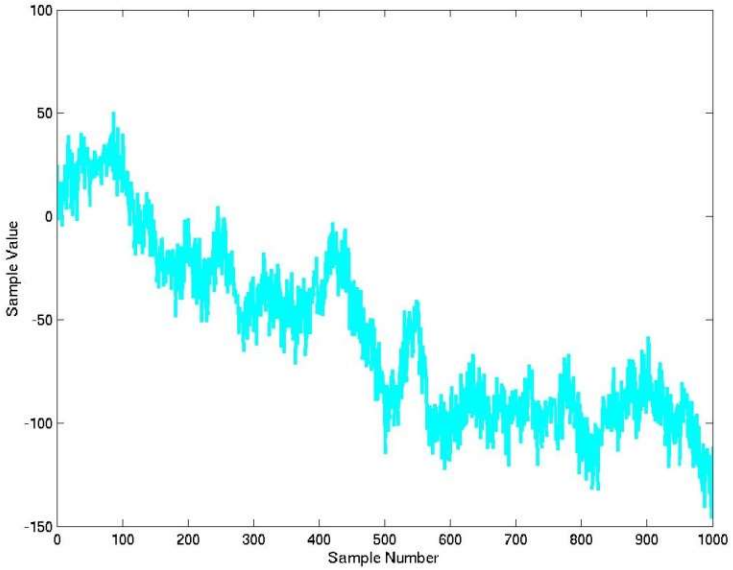
x_t is mixture Gaussian with parameters:

Means = [-4, 0, 4, 8, 12, 16, 18, 20]

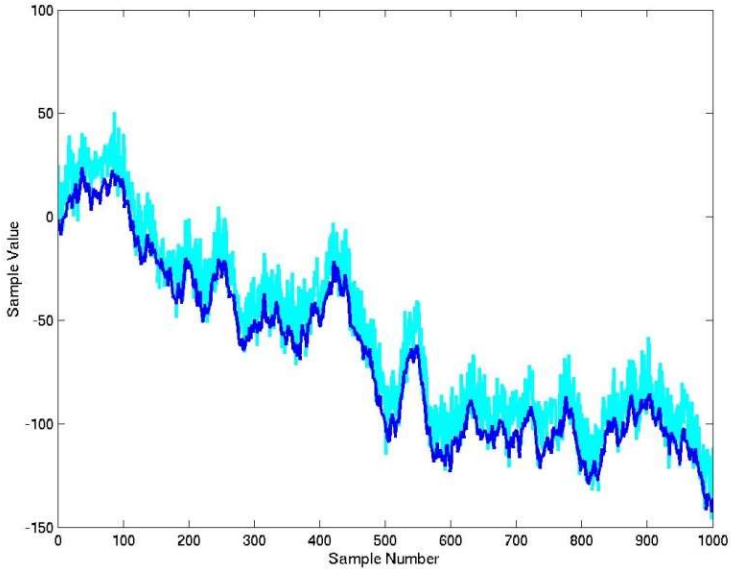
Variances = [10, 10, 10, 10, 10, 10, 10, 10]

Mixture weights = [0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125]



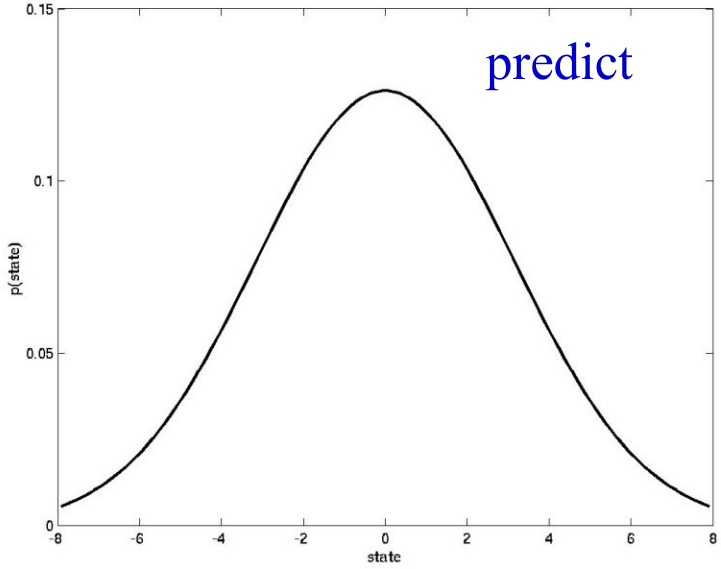


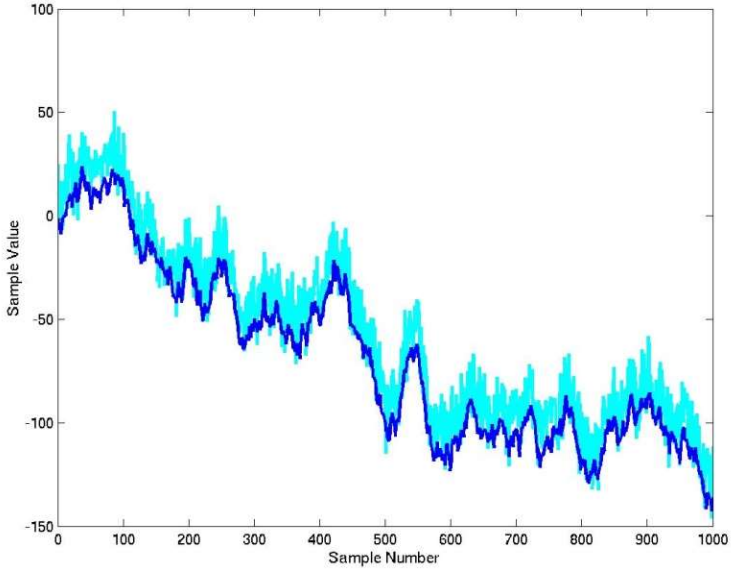
Simulation: Synthesizing data



Combined figure for more compact representation

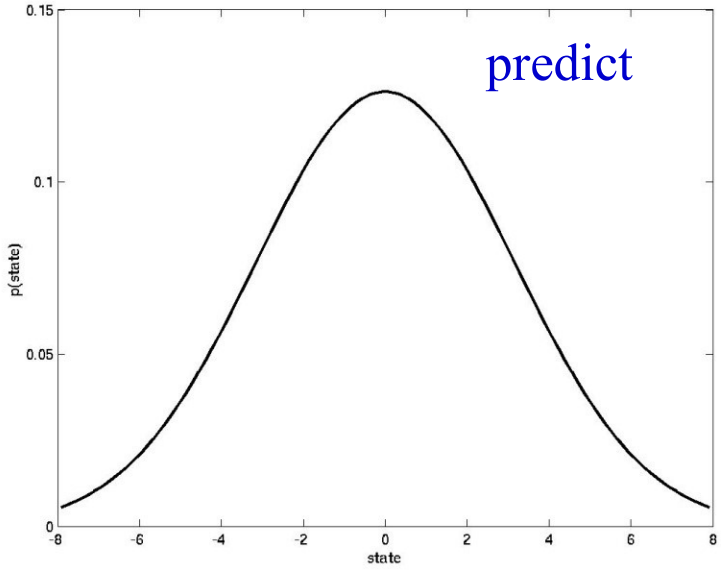
SIMULATION: TIME = 1

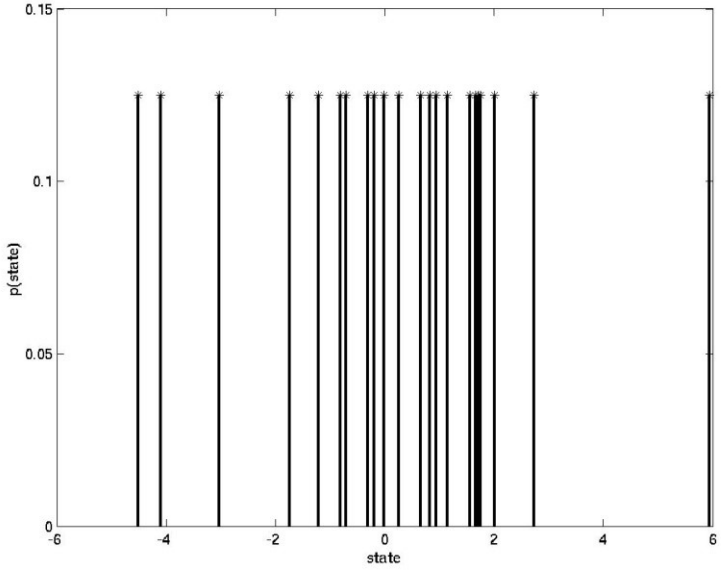


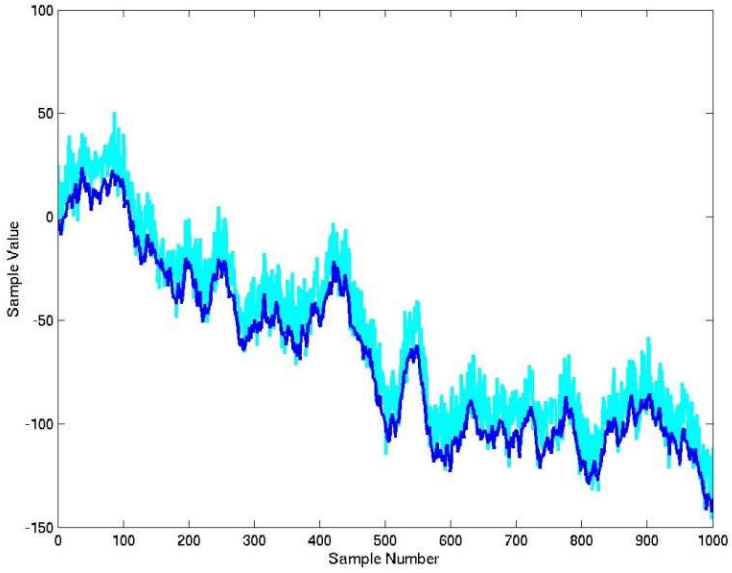


PREDICTED STATE DISTRIBUTION AT TIME = 1

SIMULATION: TIME = 1

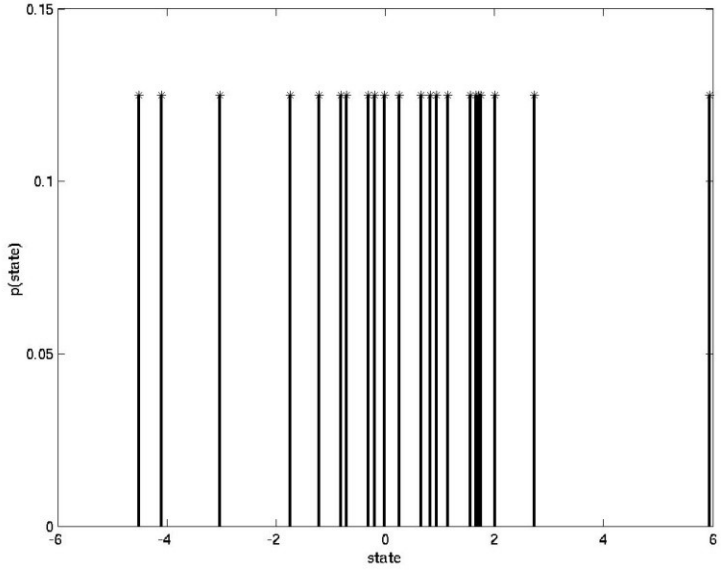


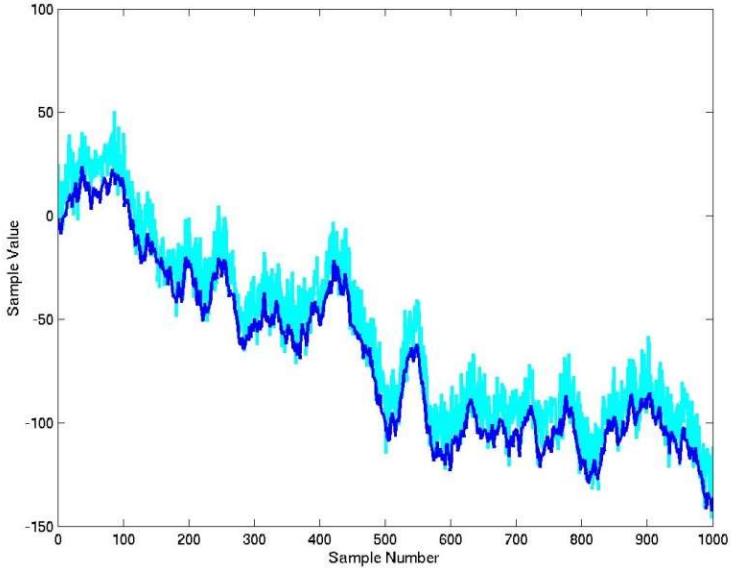




SAMPLED VERSION OF PREDICTED STATE DISTRIBUTION AT TIME = 1

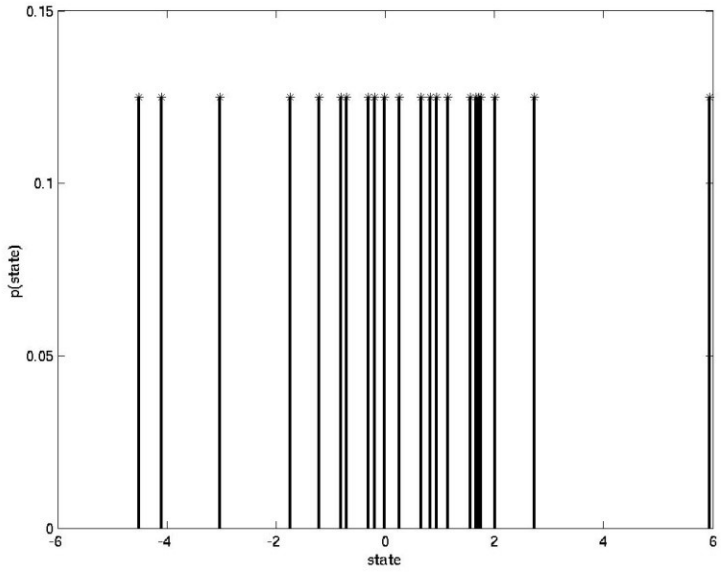
SIMULATION: TIME = 1

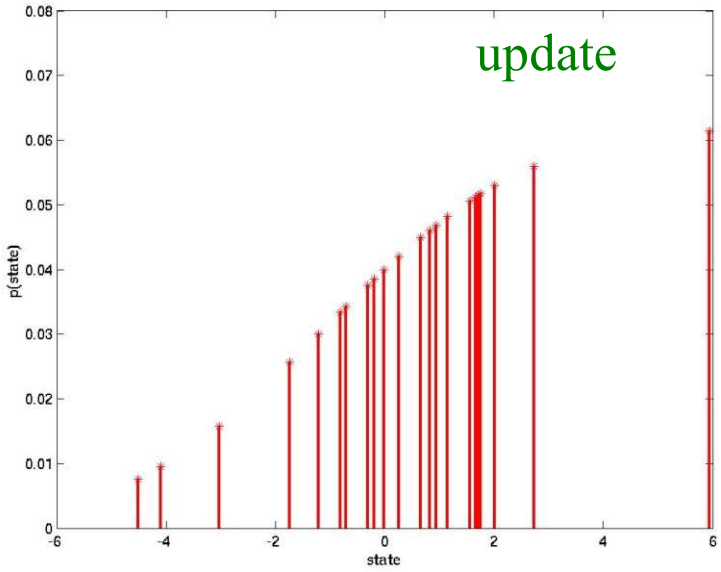


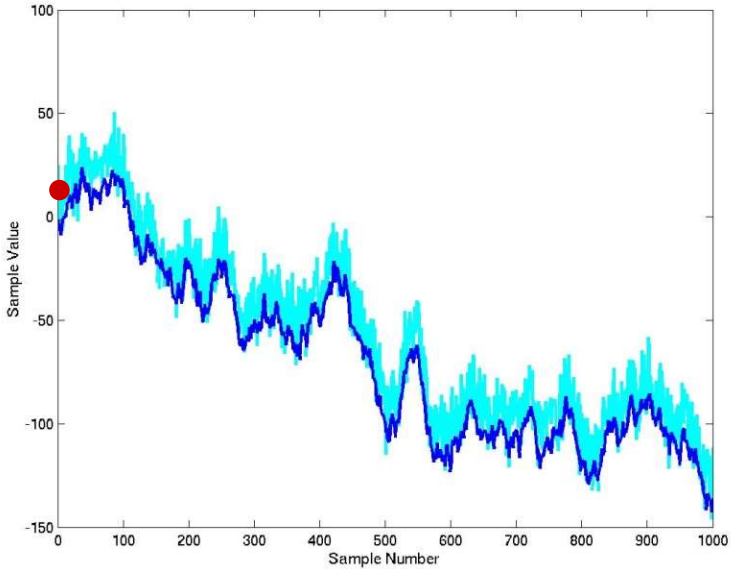


SAMPLED VERSION OF PREDICTED STATE DISTRIBUTION AT TIME = 1

SIMULATION: TIME = 1

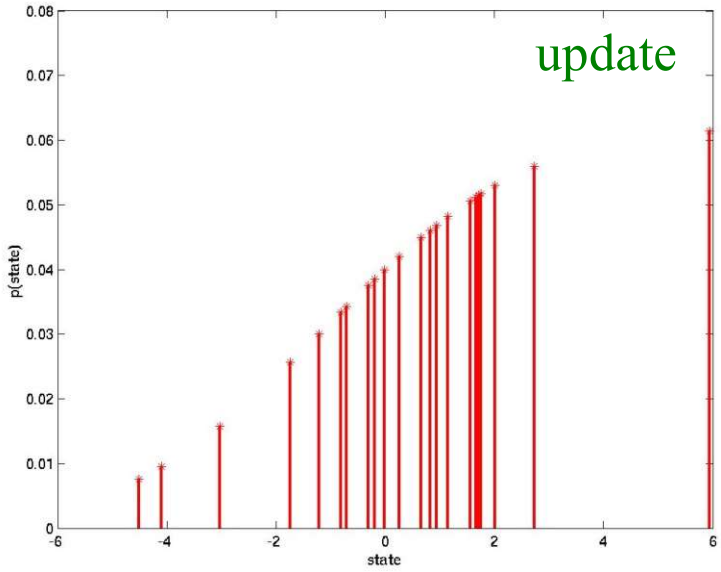


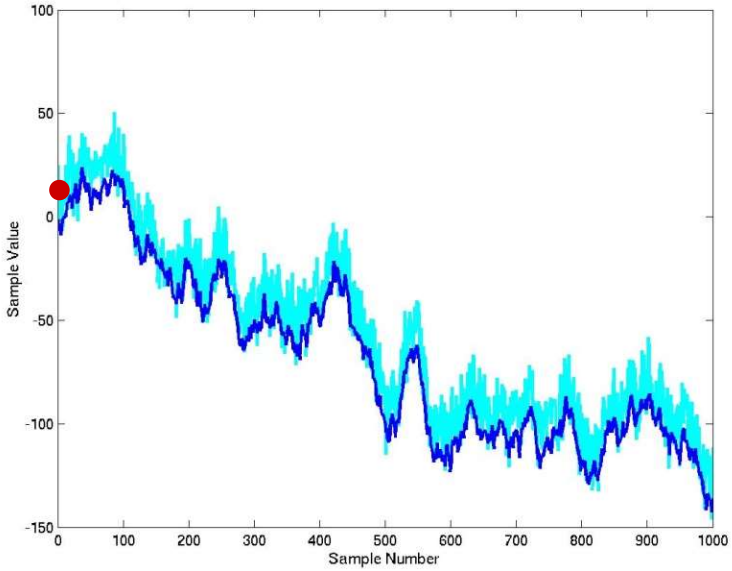


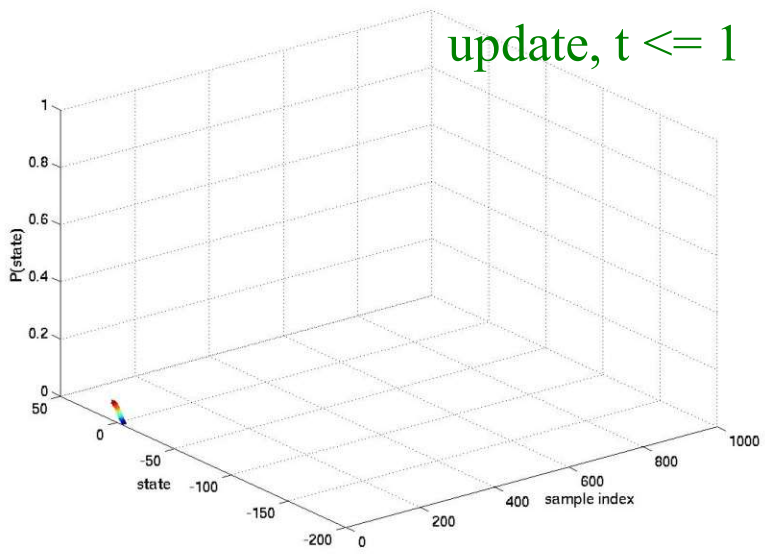


UPDATED VERSION OF
SAMPLED VERSION OF
PREDICTED STATE DISTRIBUTION
AT TIME = 1
AFTER SEEING FIRST OBSERVATION

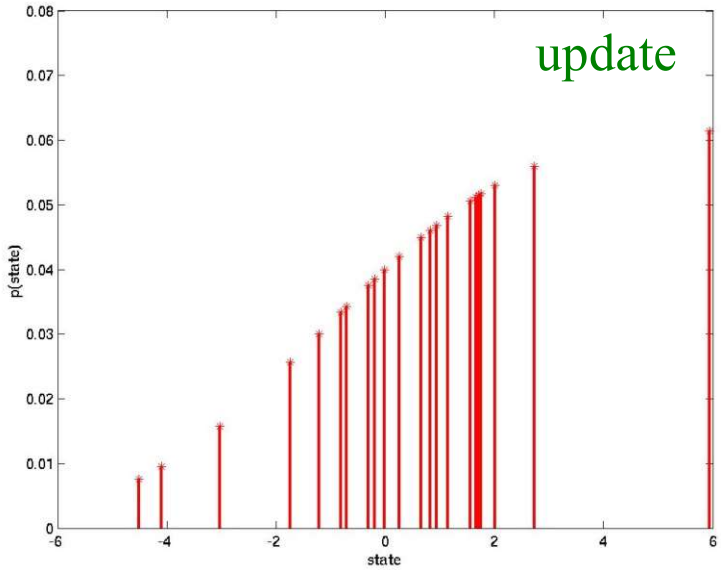
SIMULATION: TIME = 1

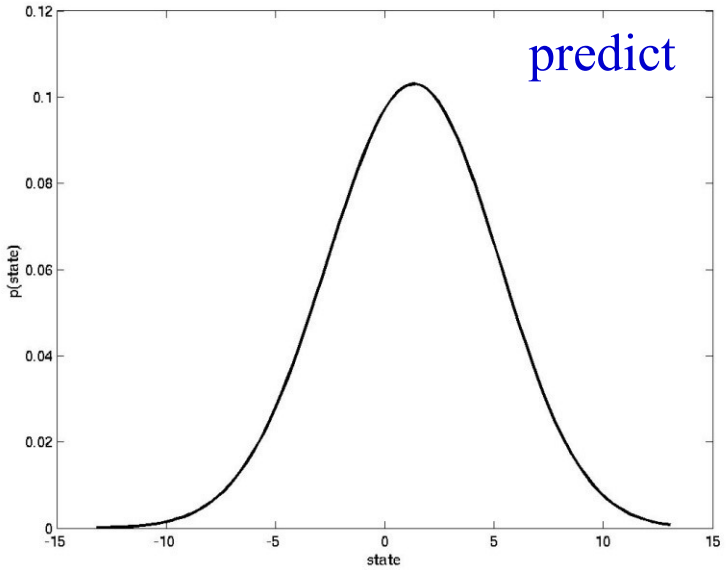


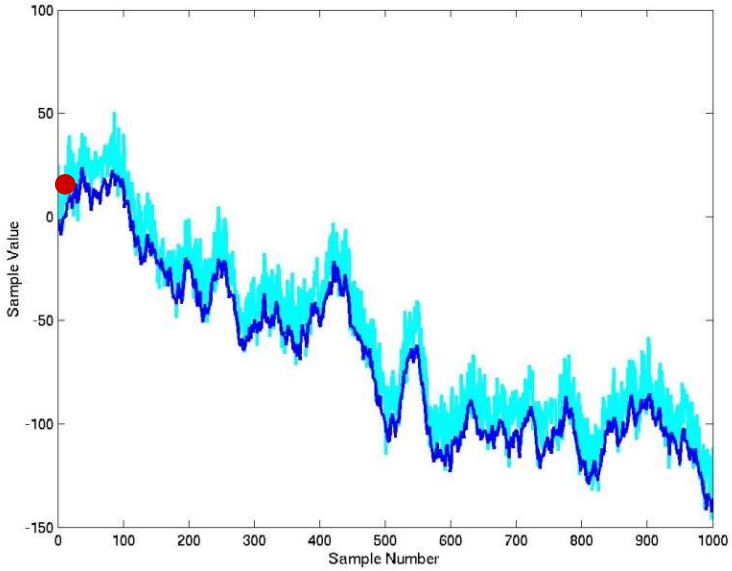


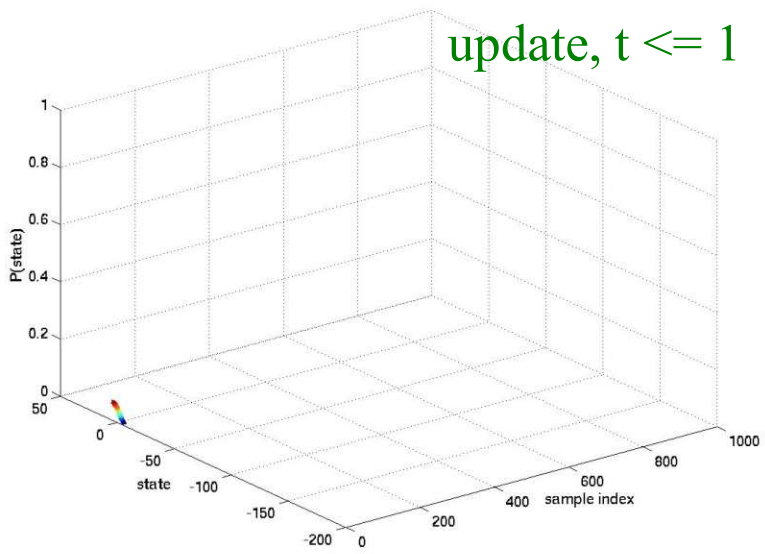


SIMULATION: TIME = 2

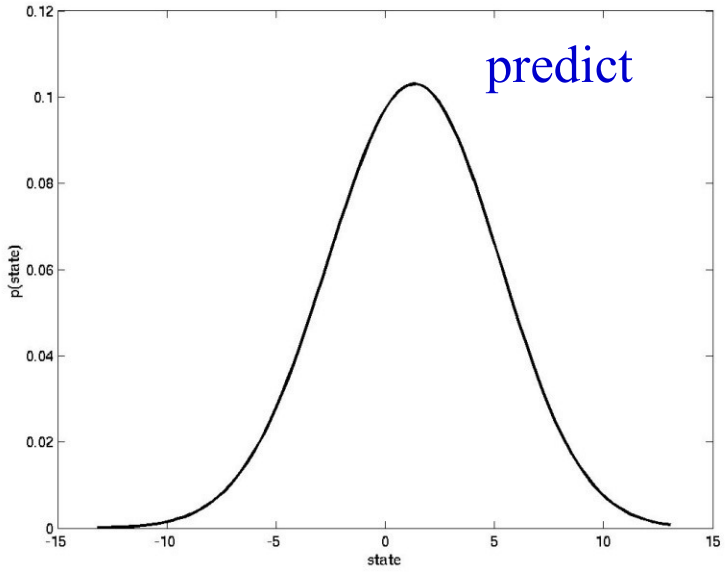


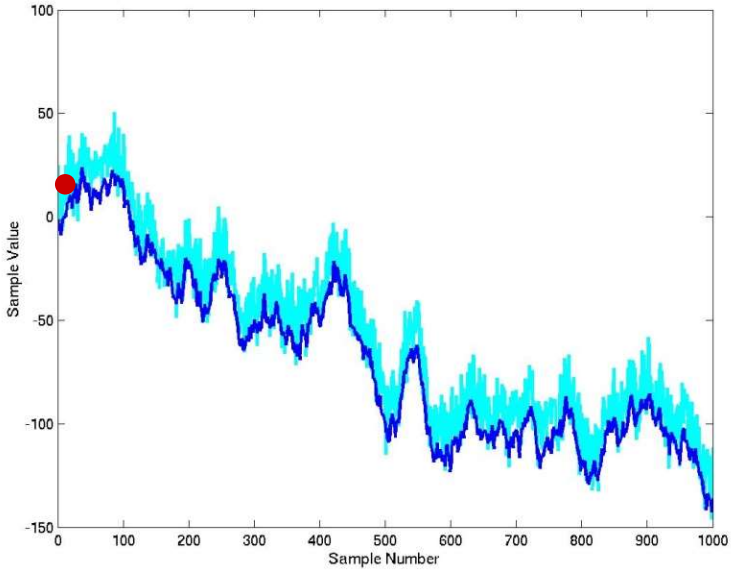


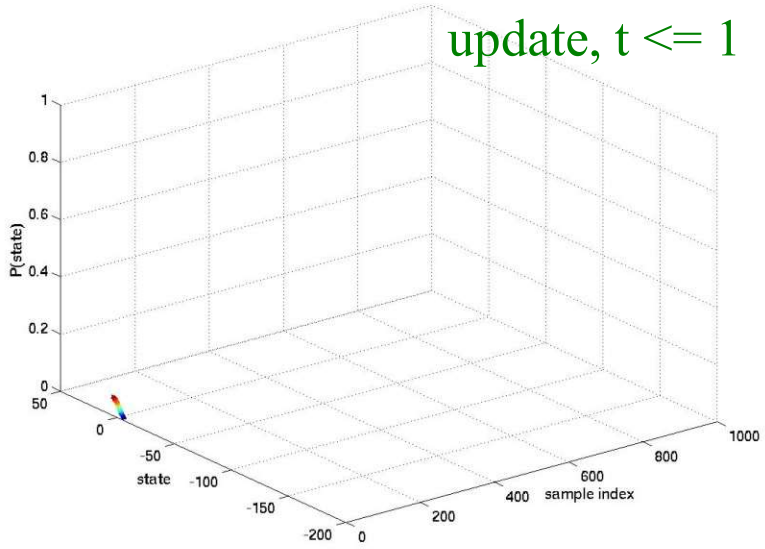




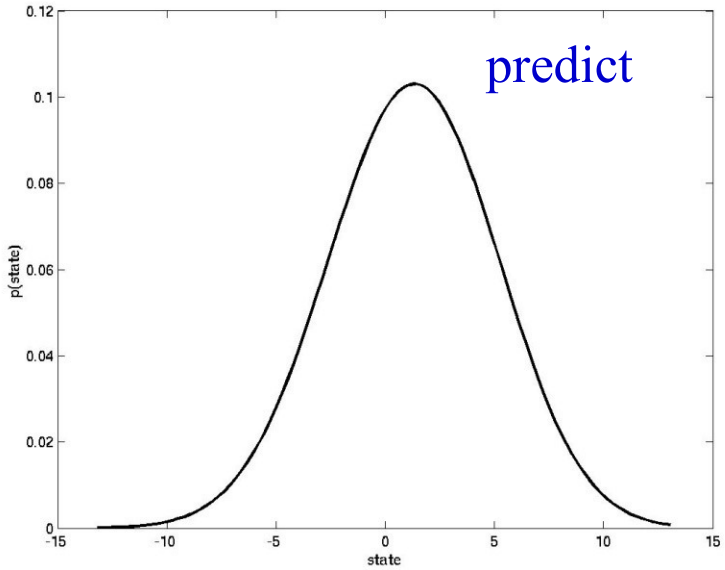
SIMULATION: TIME = 2

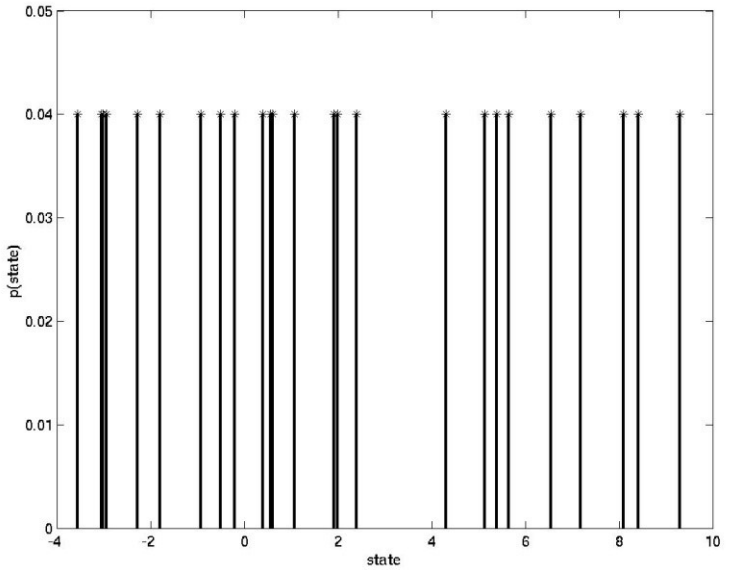


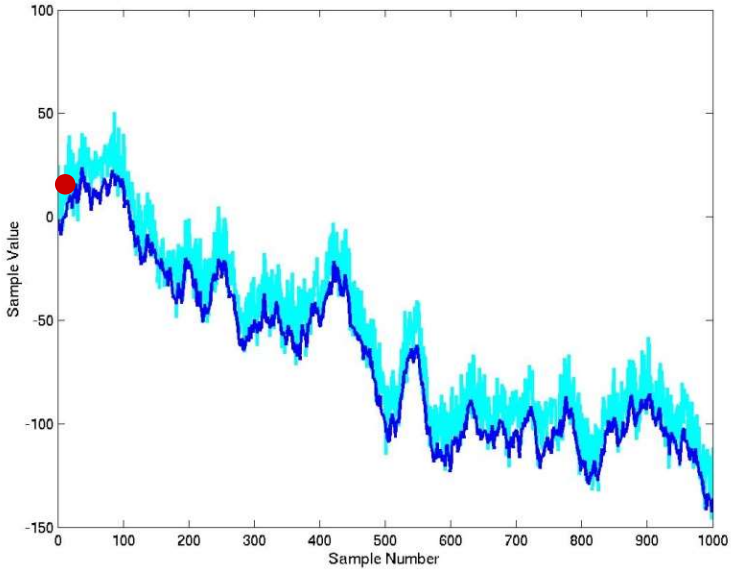


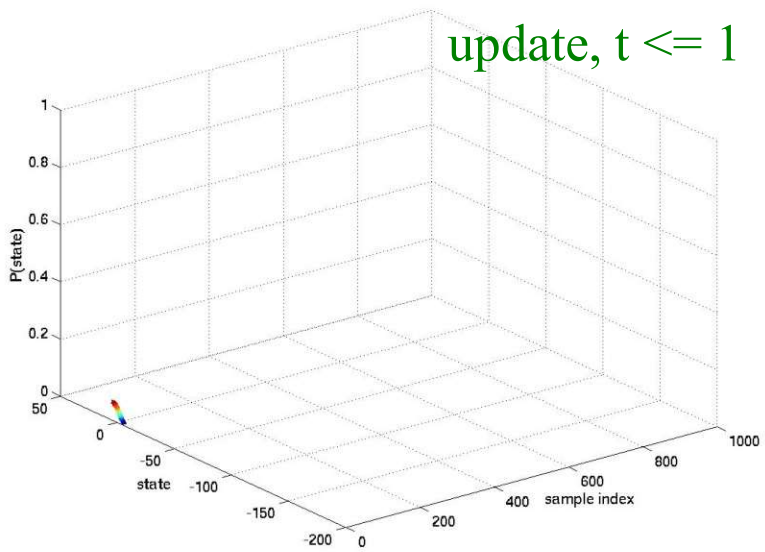


SIMULATION: TIME = 2

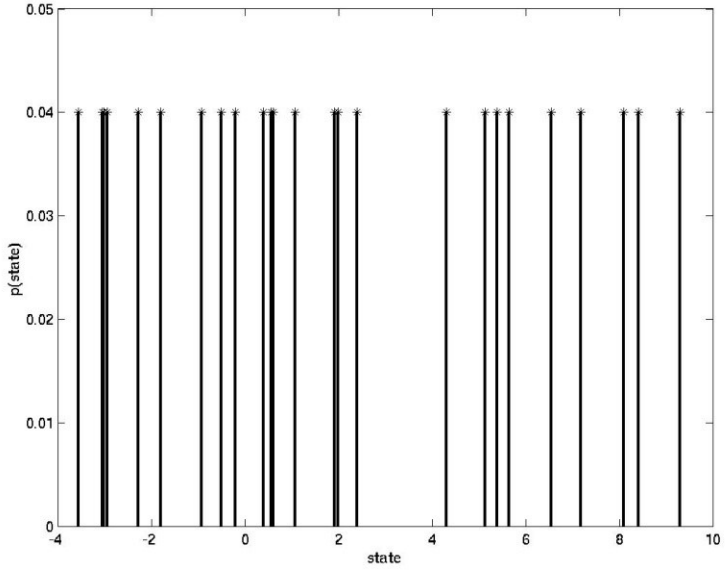


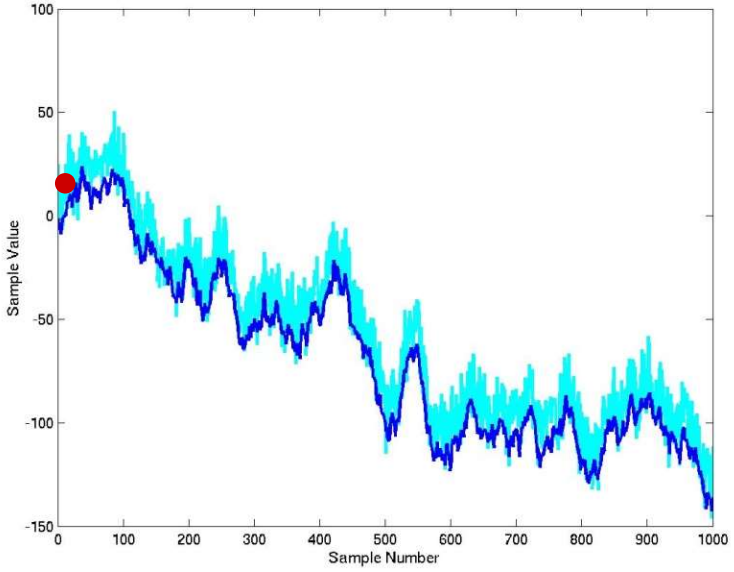




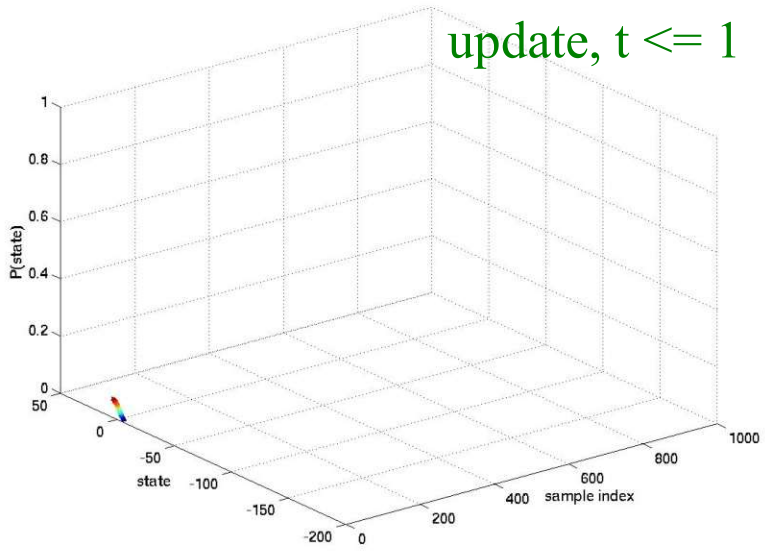


SIMULATION: TIME = 2

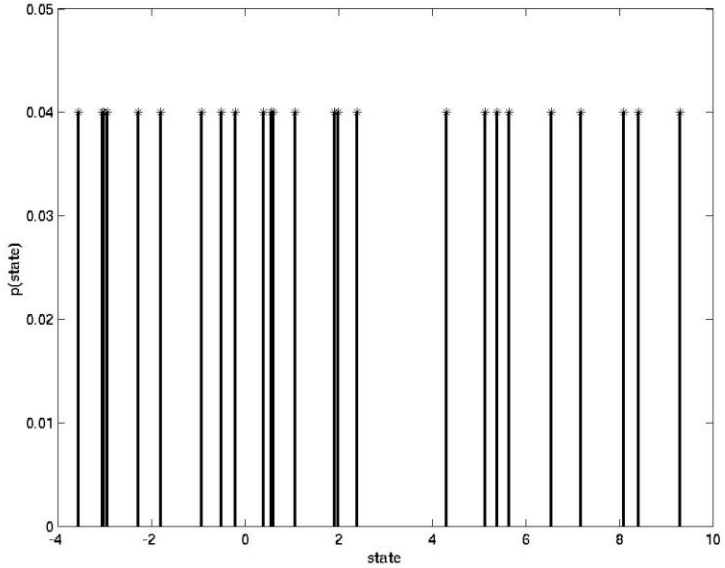


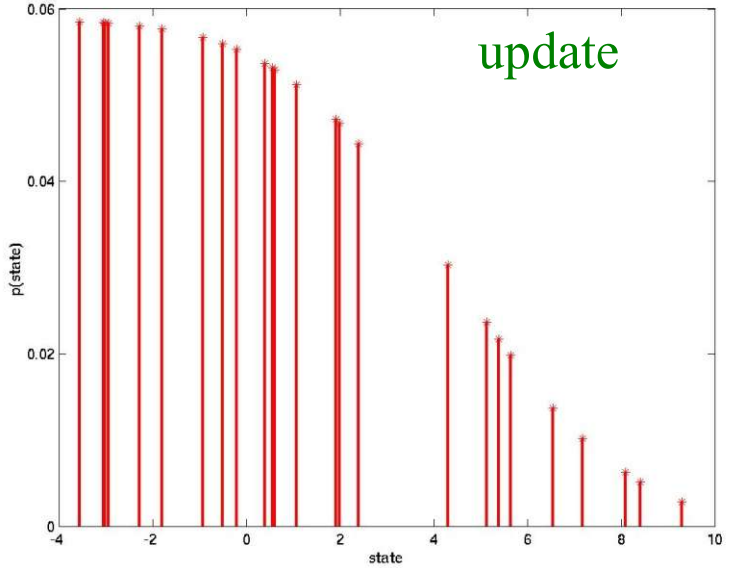


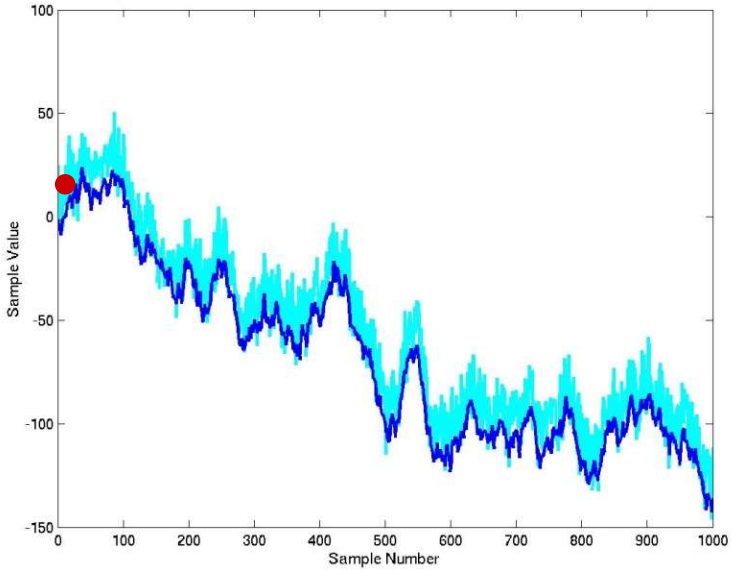
update, t <= 1

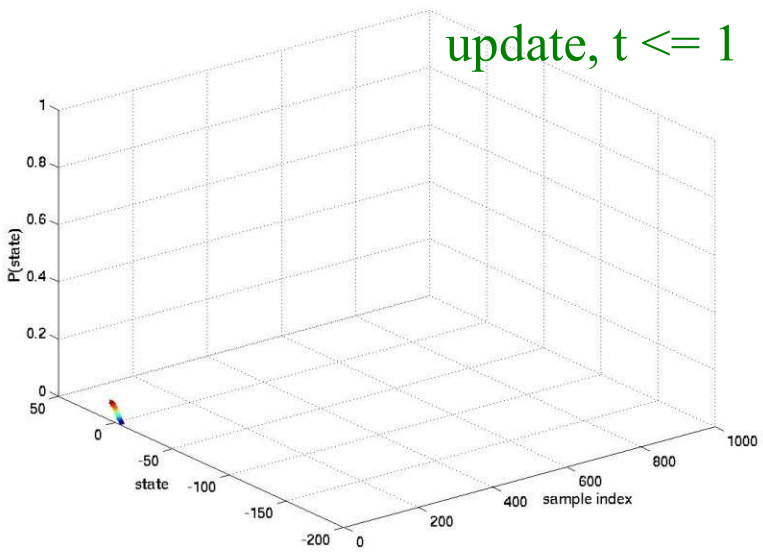


SIMULATION: TIME = 2

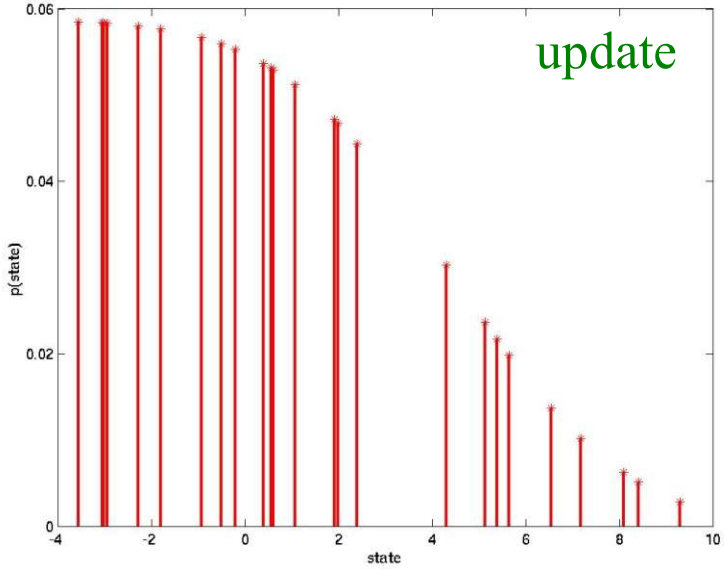


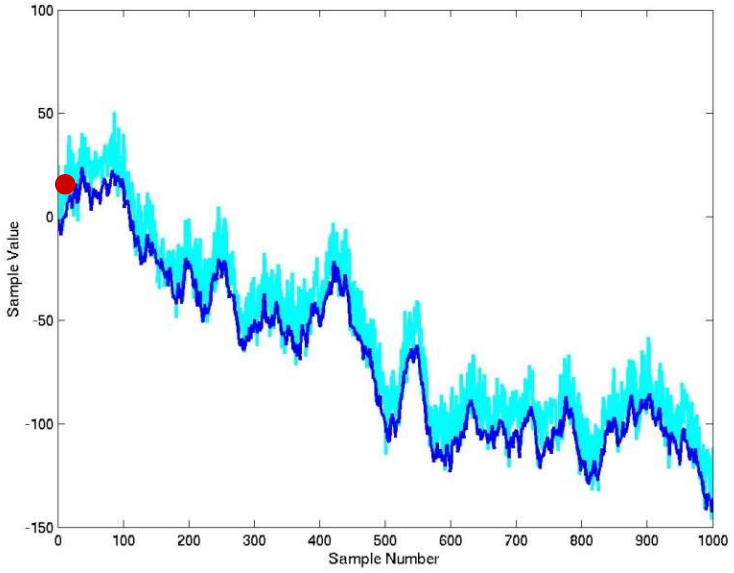


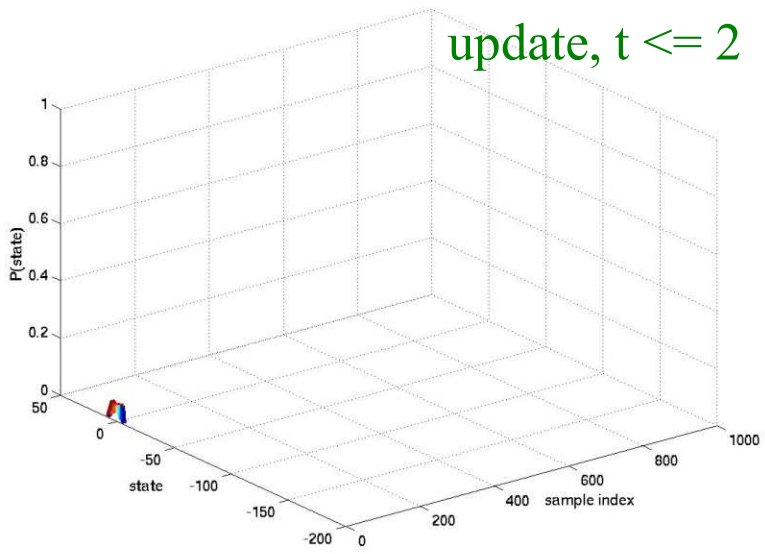




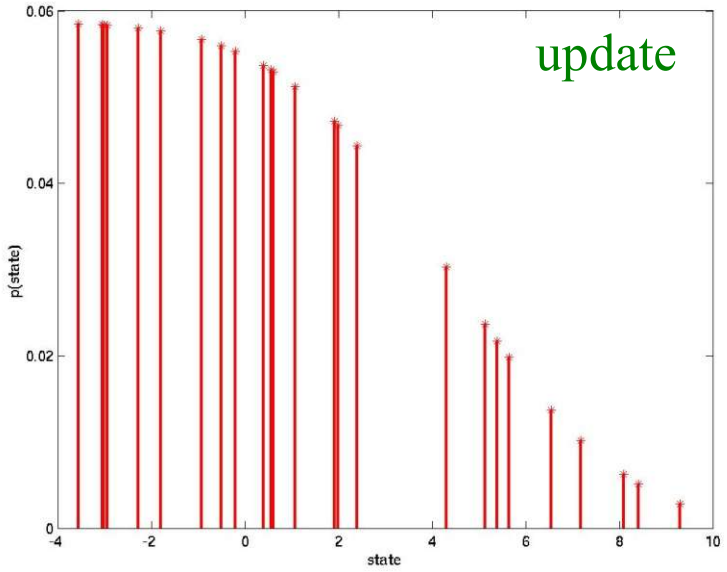
SIMULATION: TIME = 2

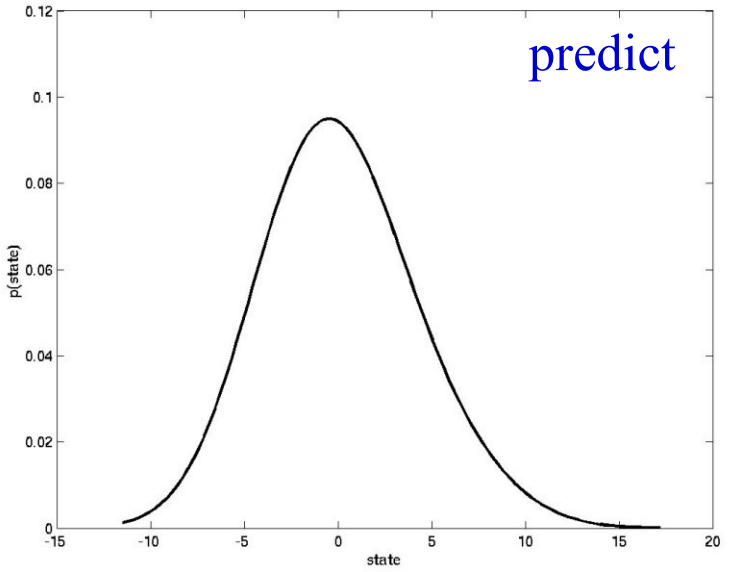


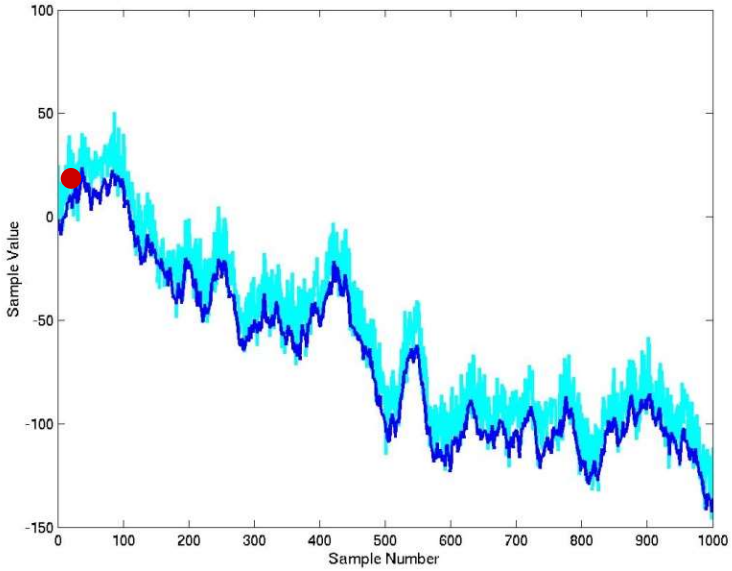


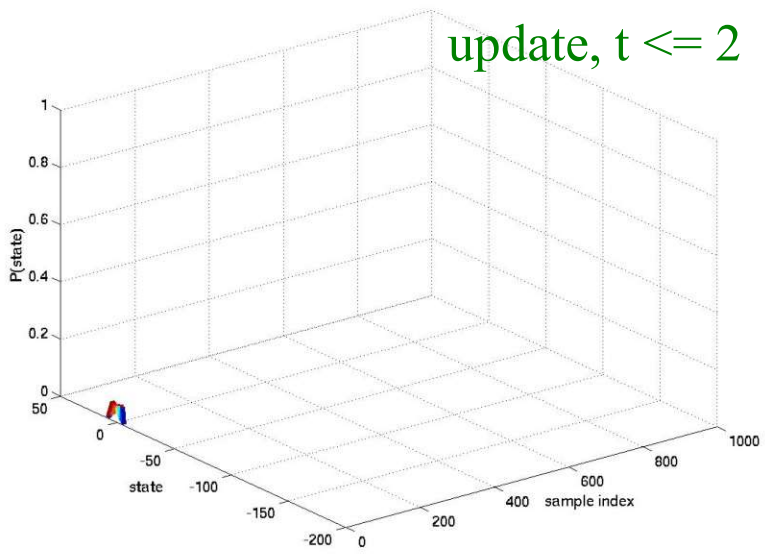


SIMULATION: TIME = 3

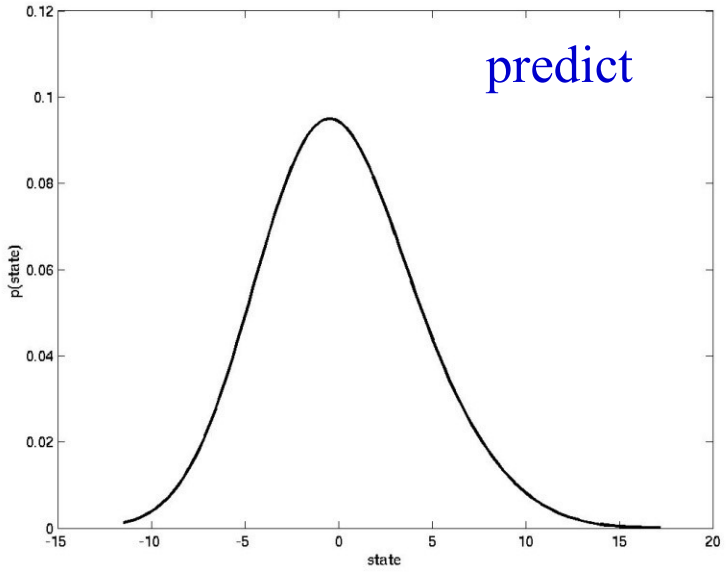


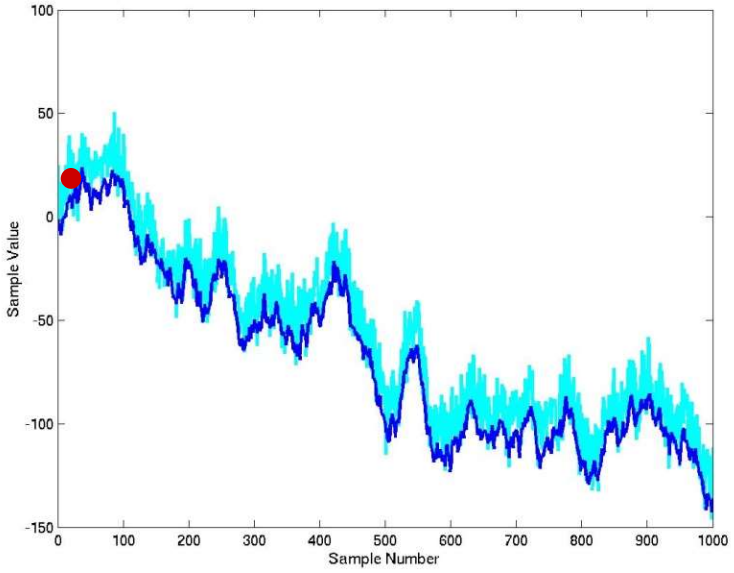




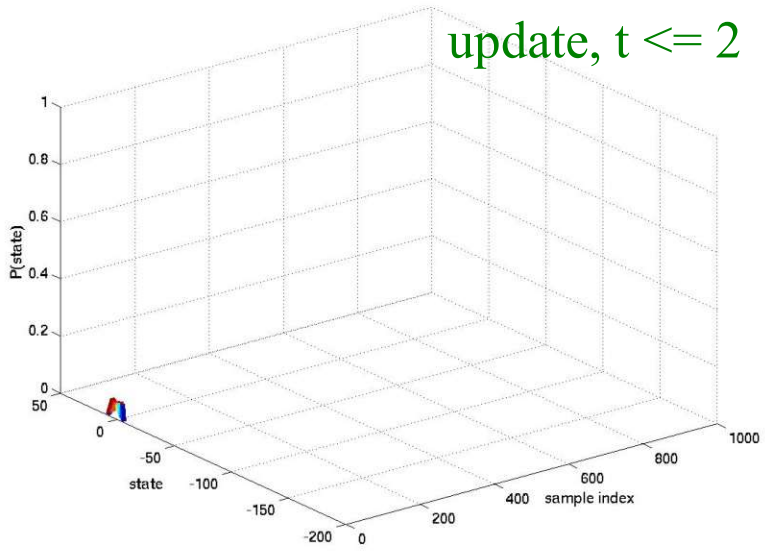


SIMULATION: TIME = 3

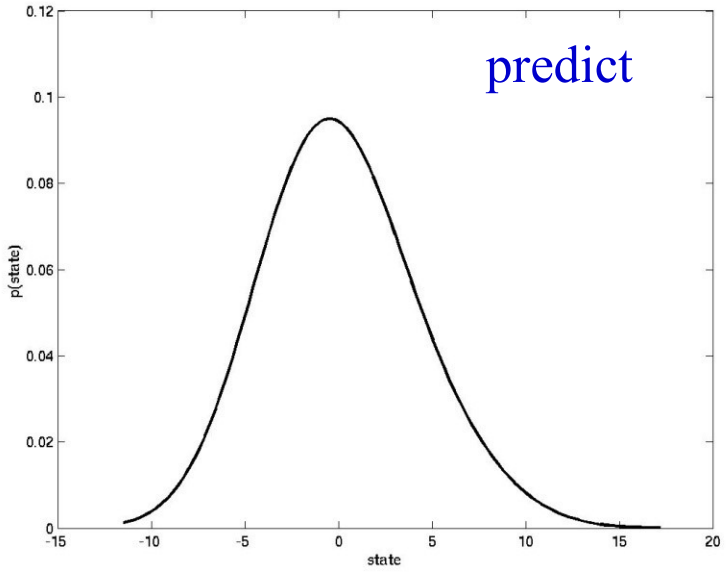


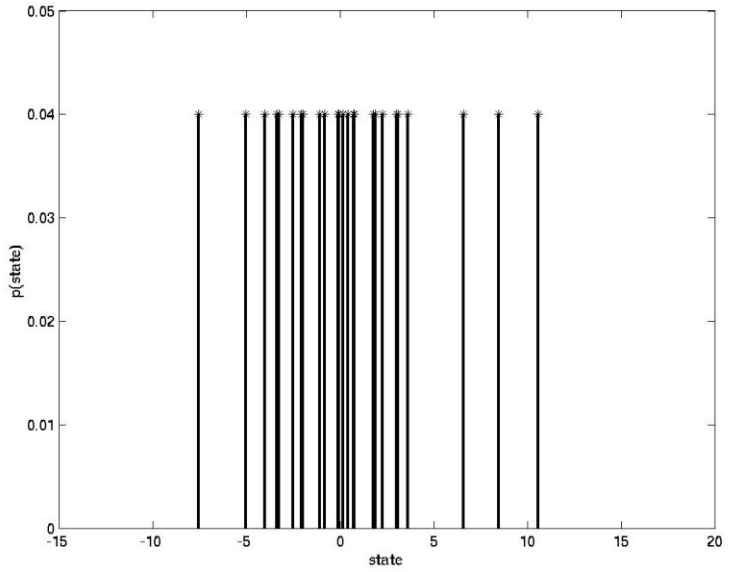


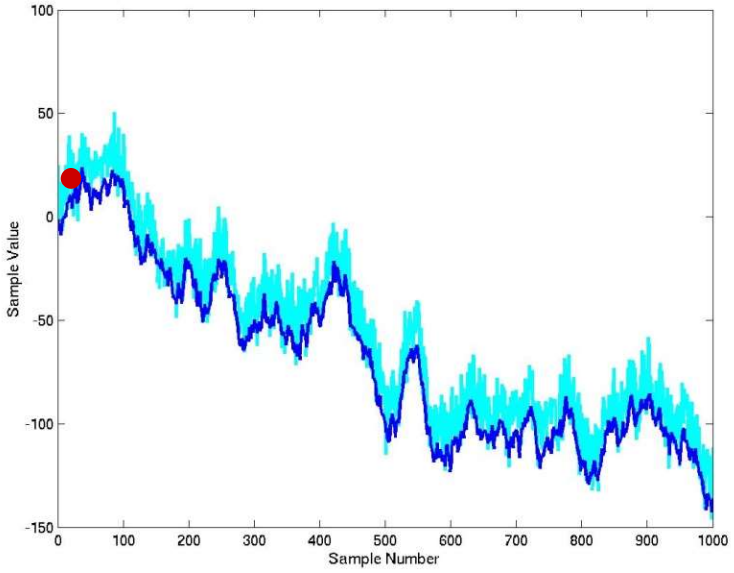
update, t <= 2

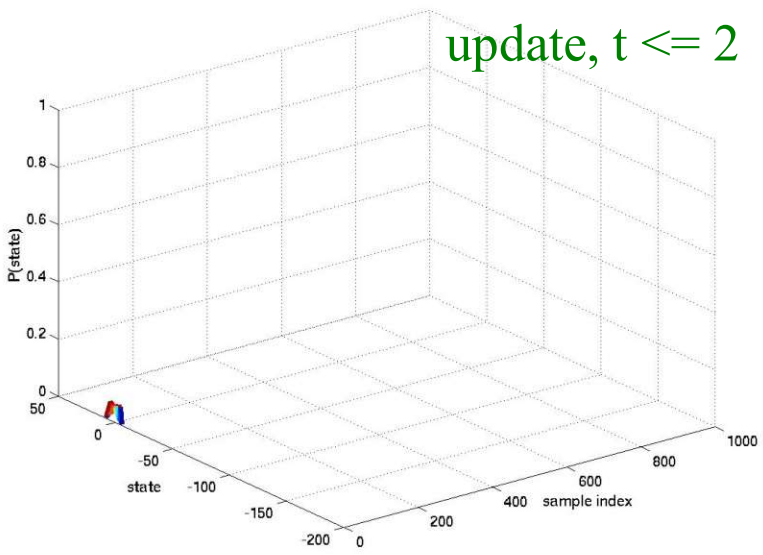


SIMULATION: TIME = 3

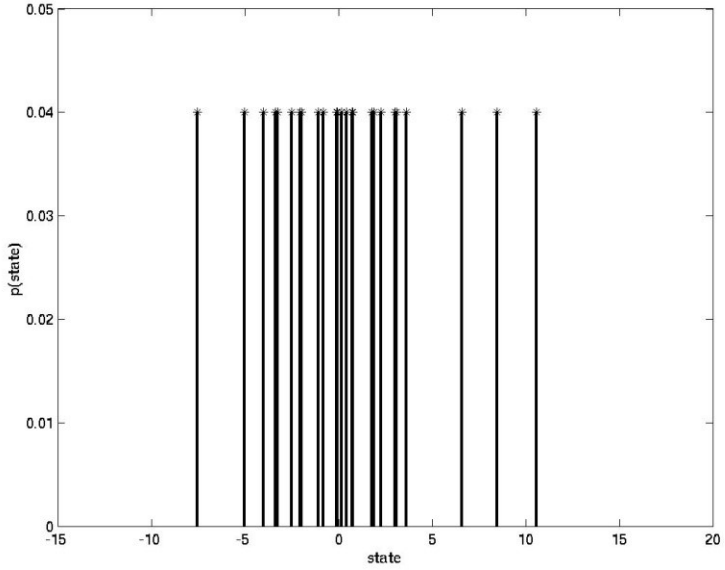


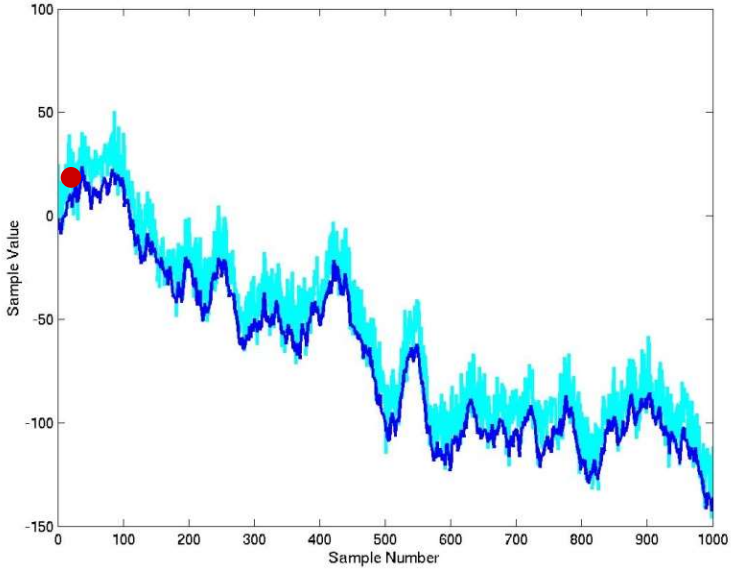




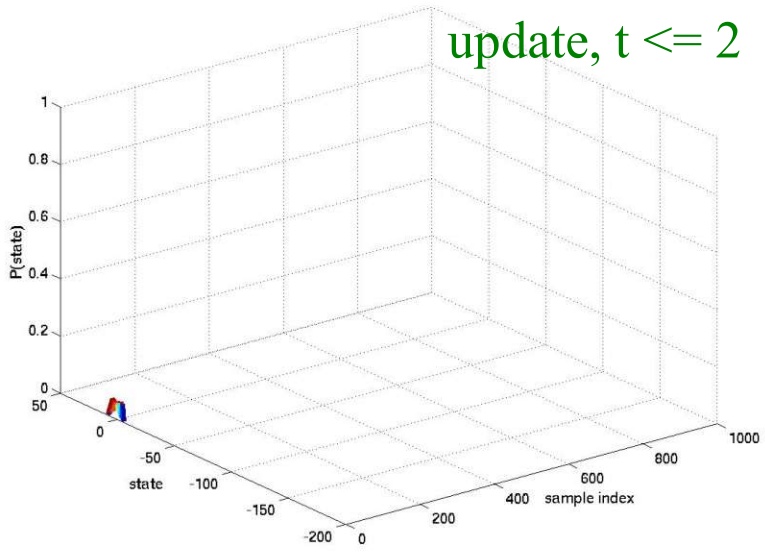


SIMULATION: TIME = 3

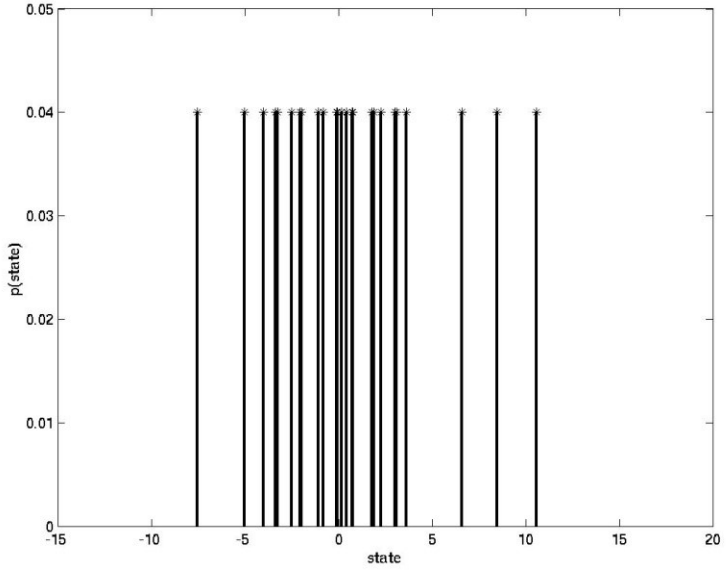


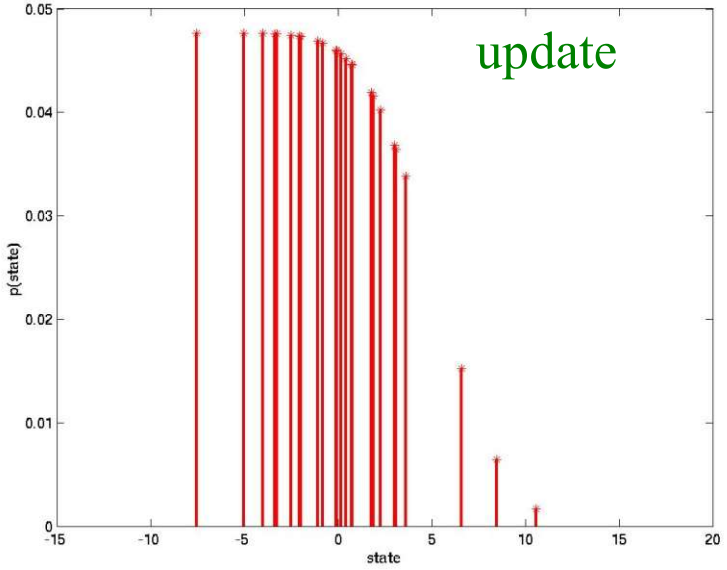


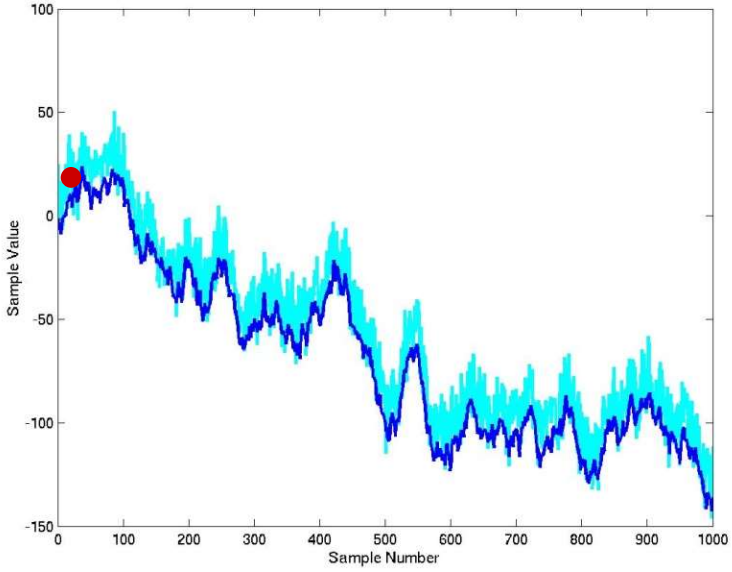
update, t <= 2

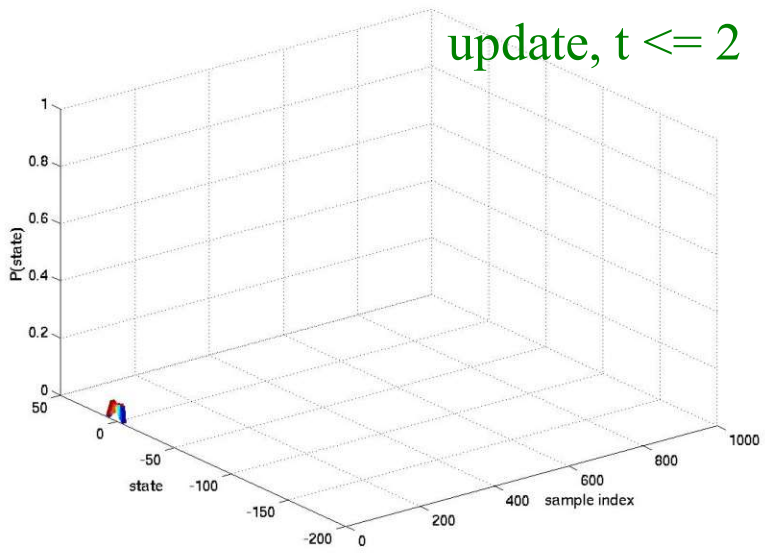


SIMULATION: TIME = 3

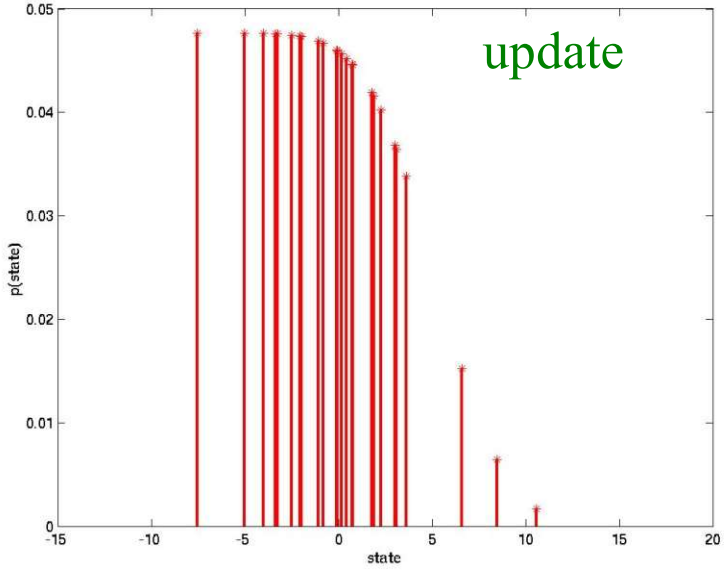




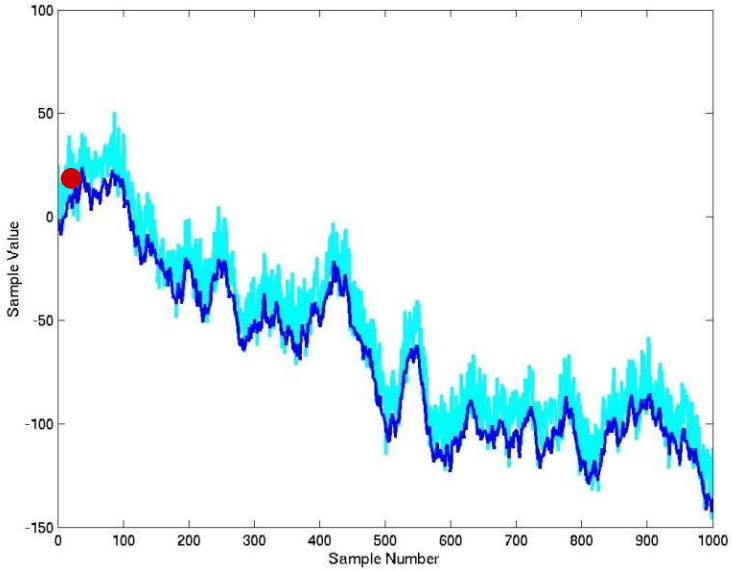


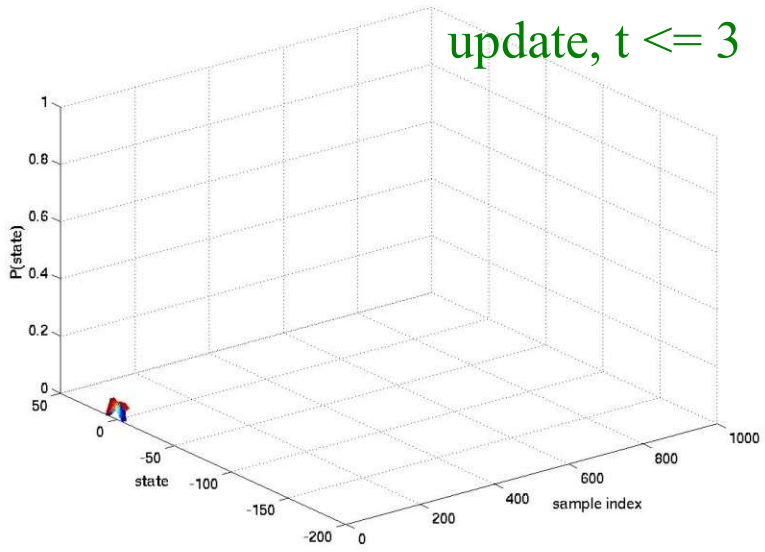


SIMULATION: TIME = 3

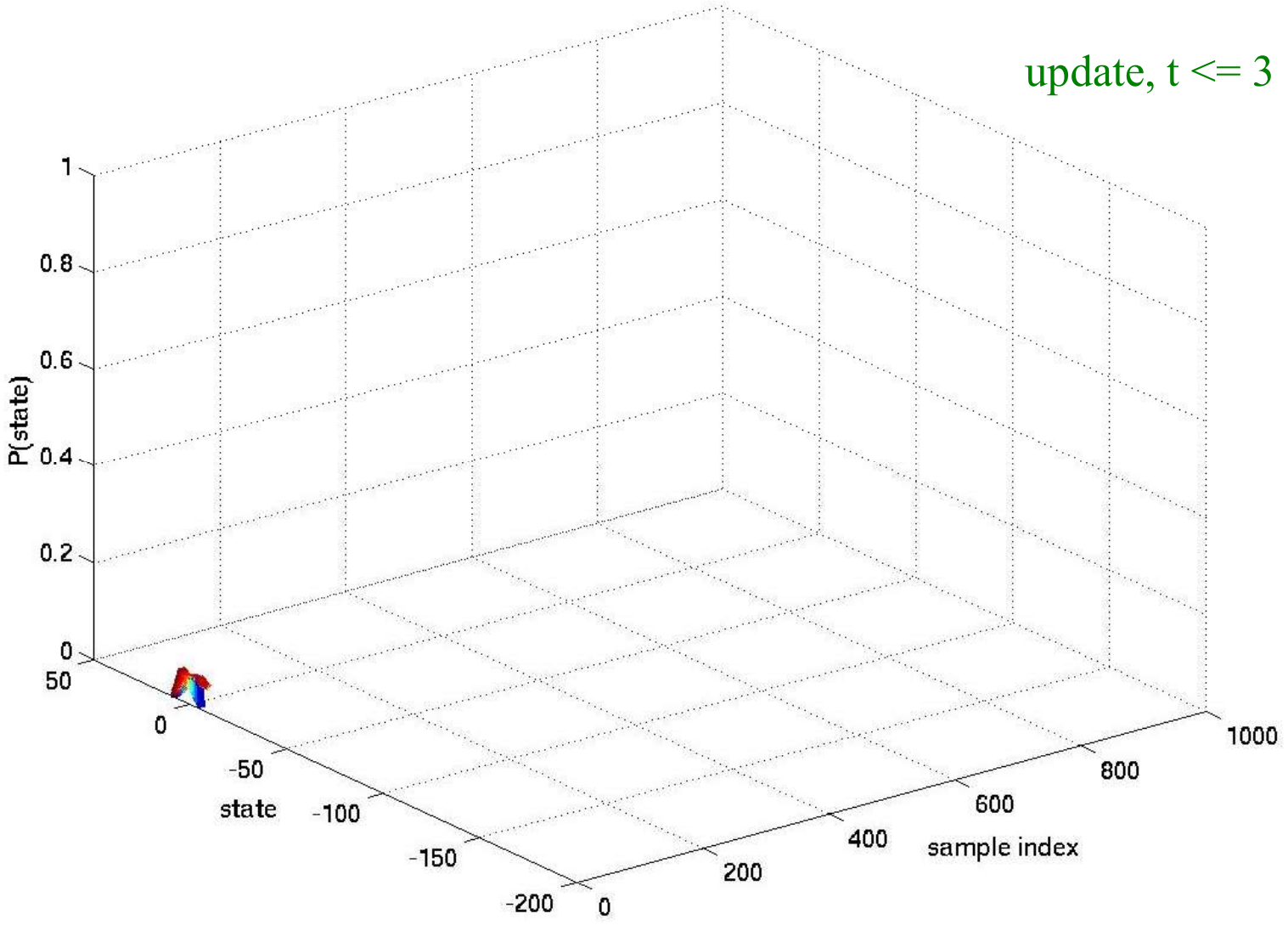


The figure below shows the contour of the updated state probabilities for all time instants until the current instant

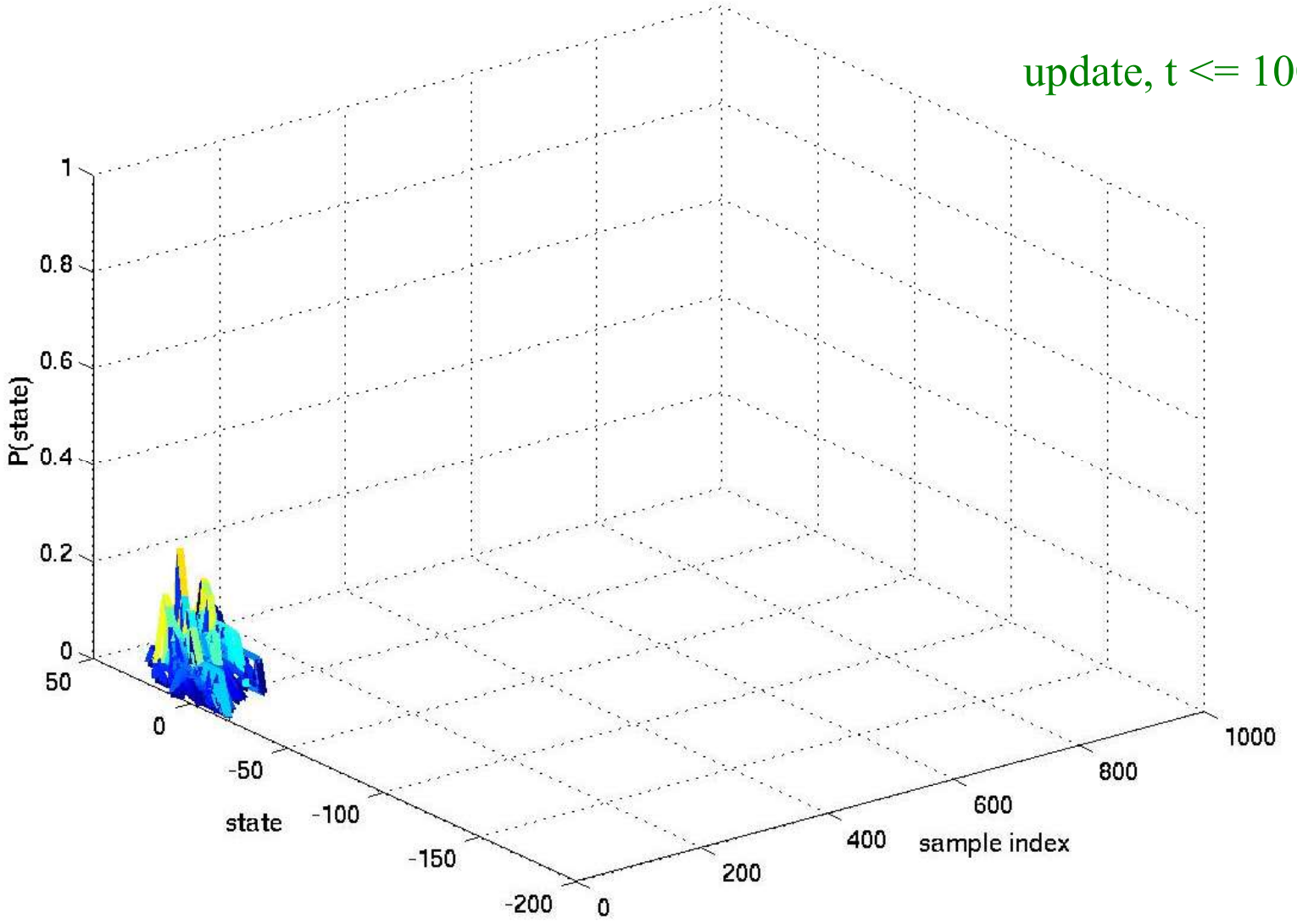




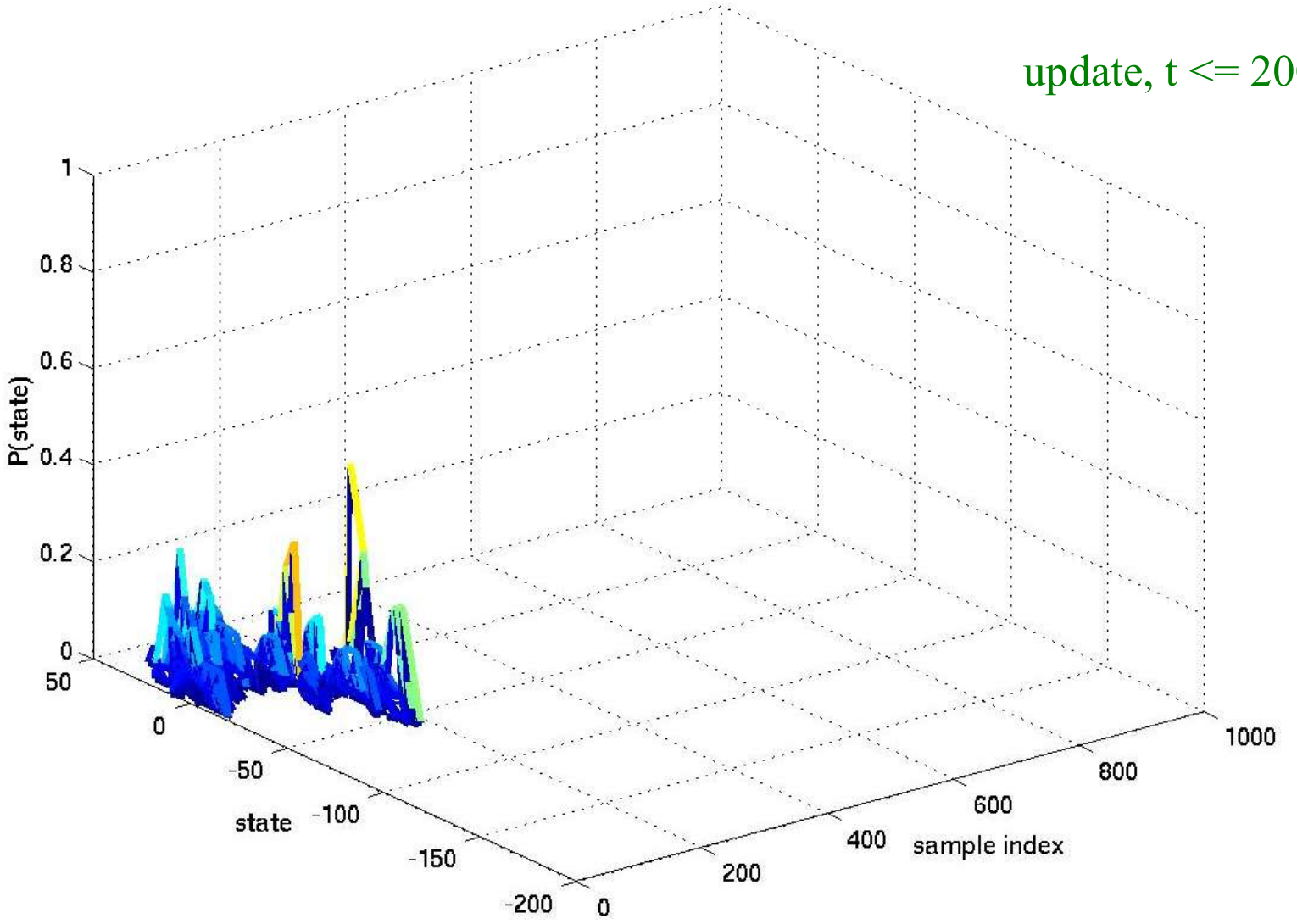
Simulation: Updated Probs Until



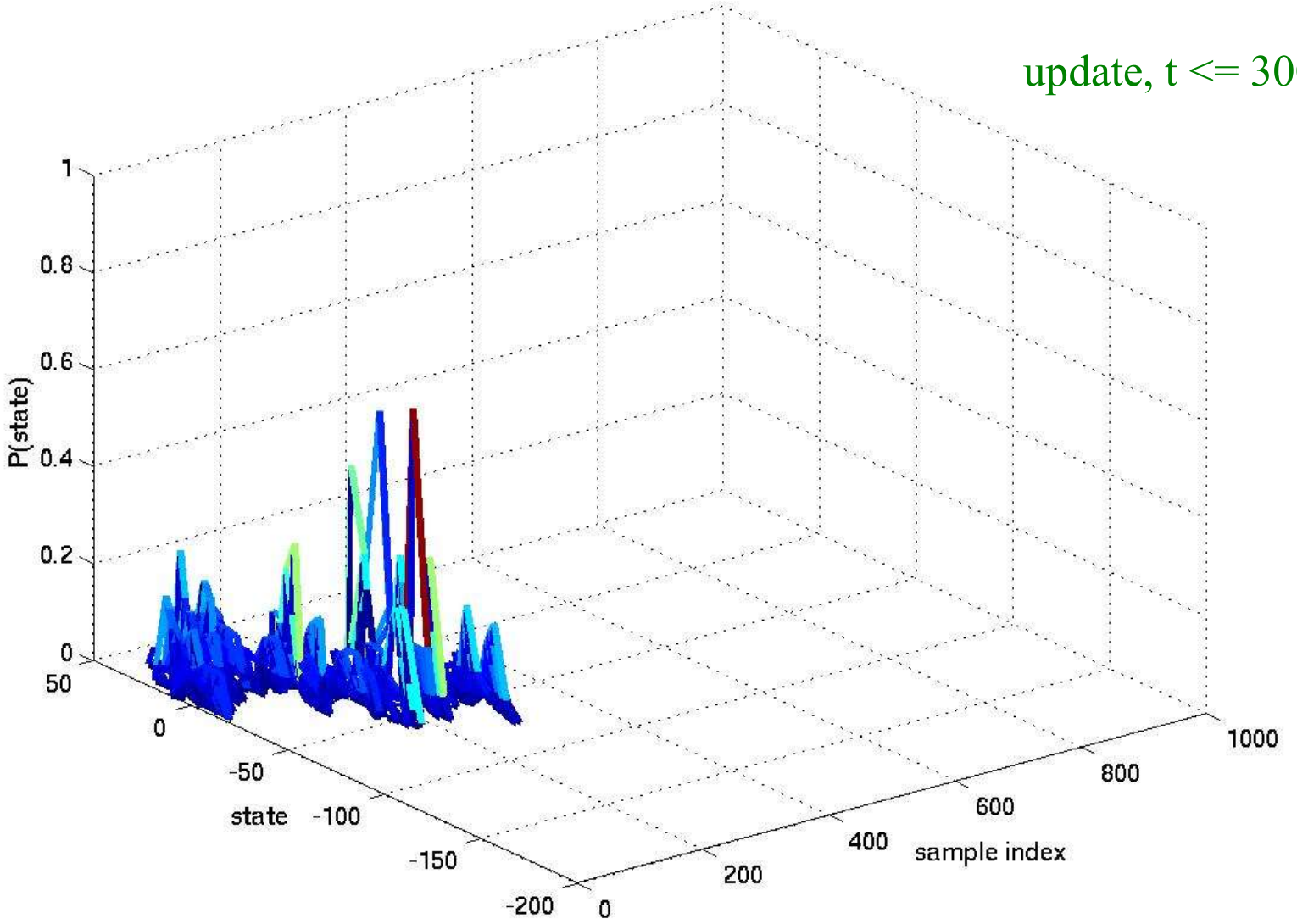
Simulation: Updated Probs Until T=100



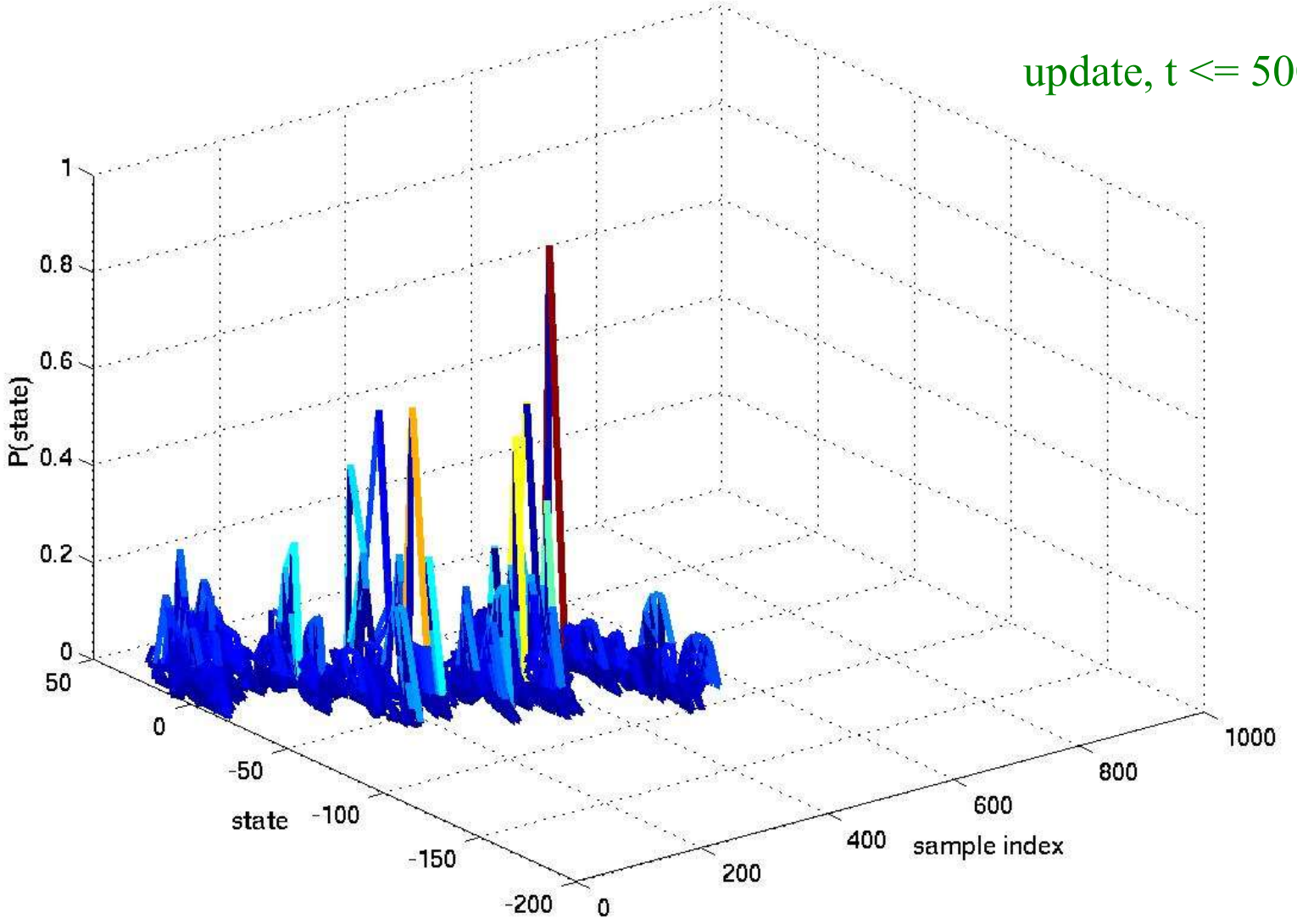
Simulation: Updated Probs Until T=200



Simulation: Updated Probs Until T=300

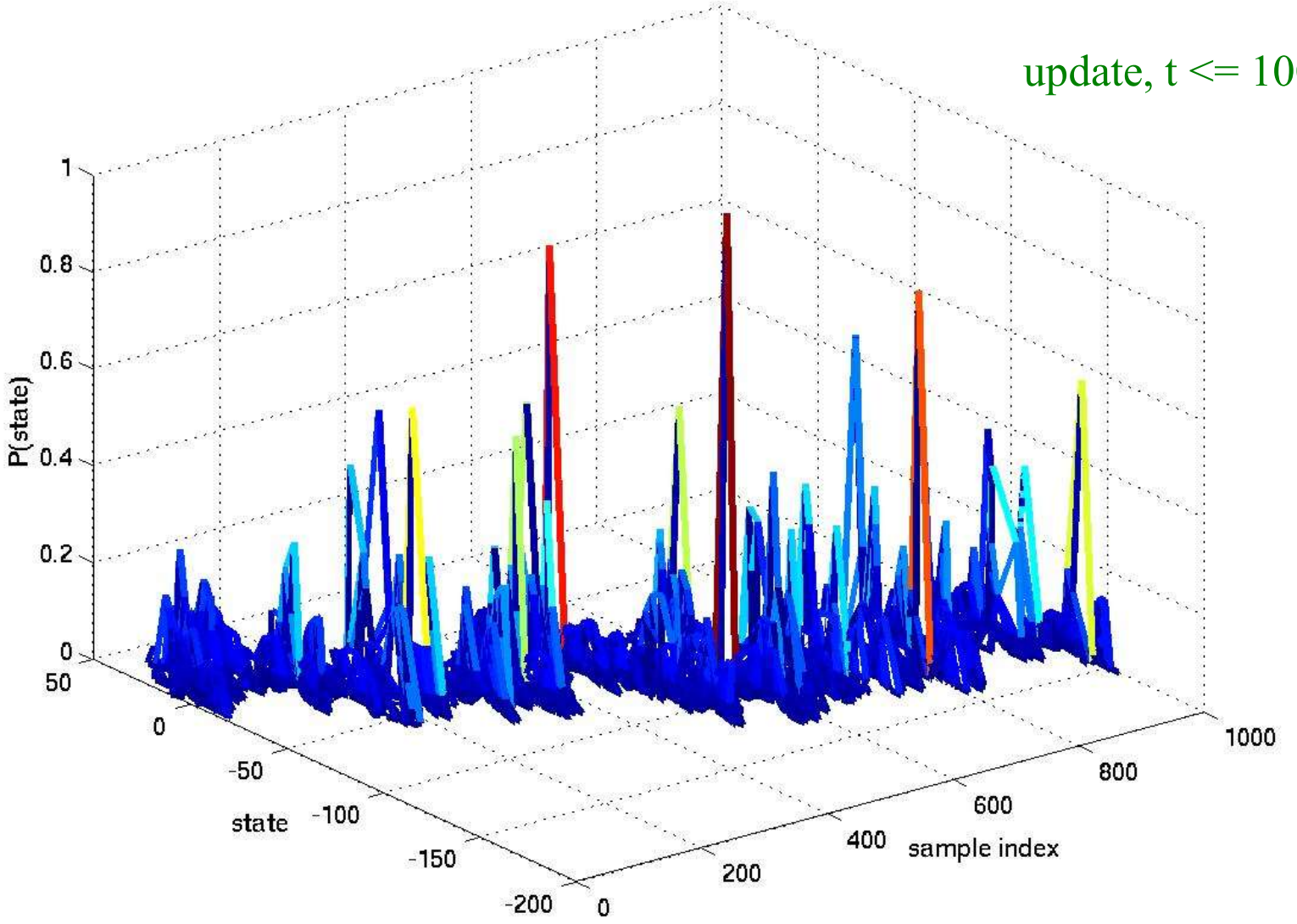


Simulation: Updated Probs Until T=500

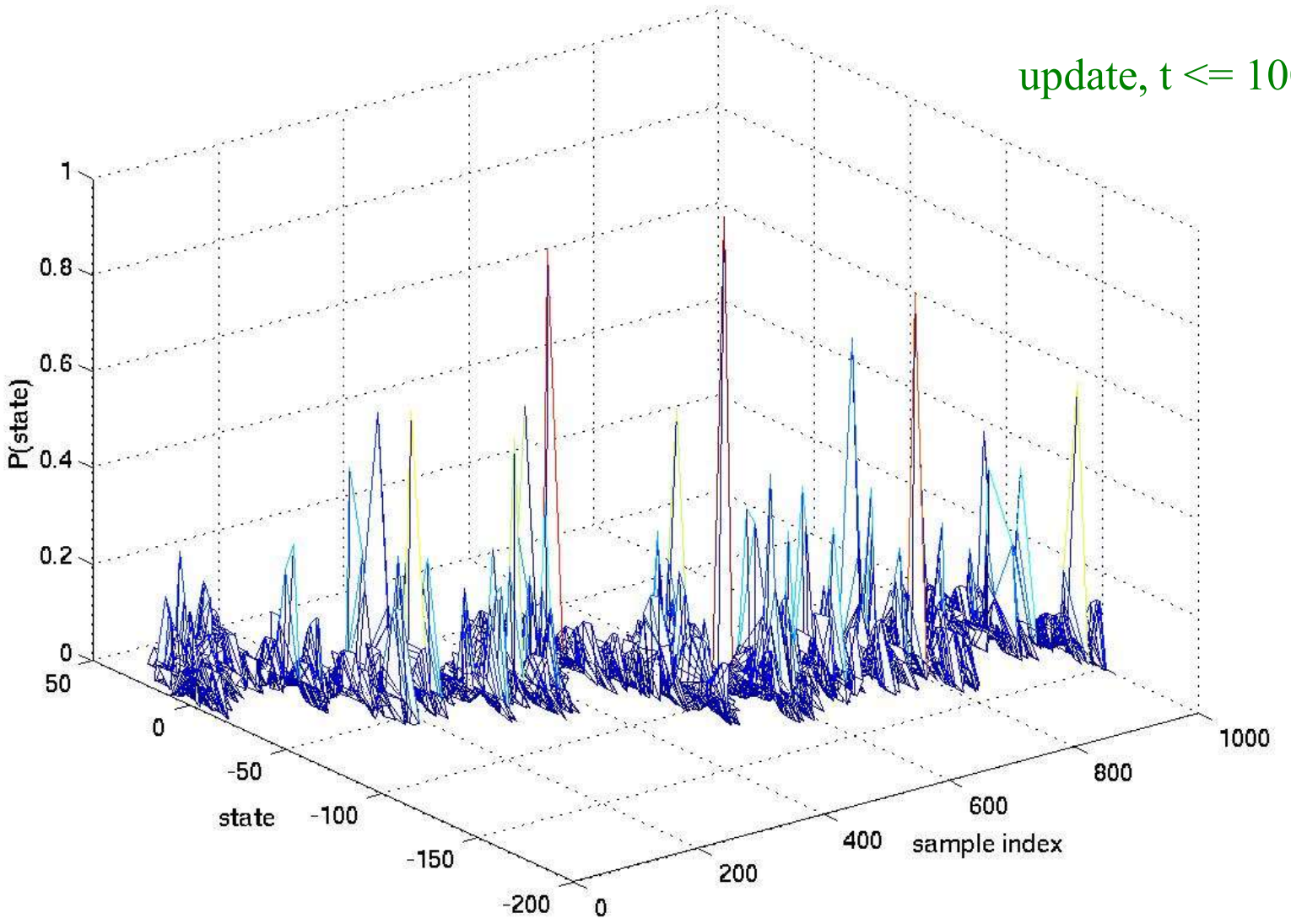


Simulation: Updated Probs Until T=1000

update, t <= 1000

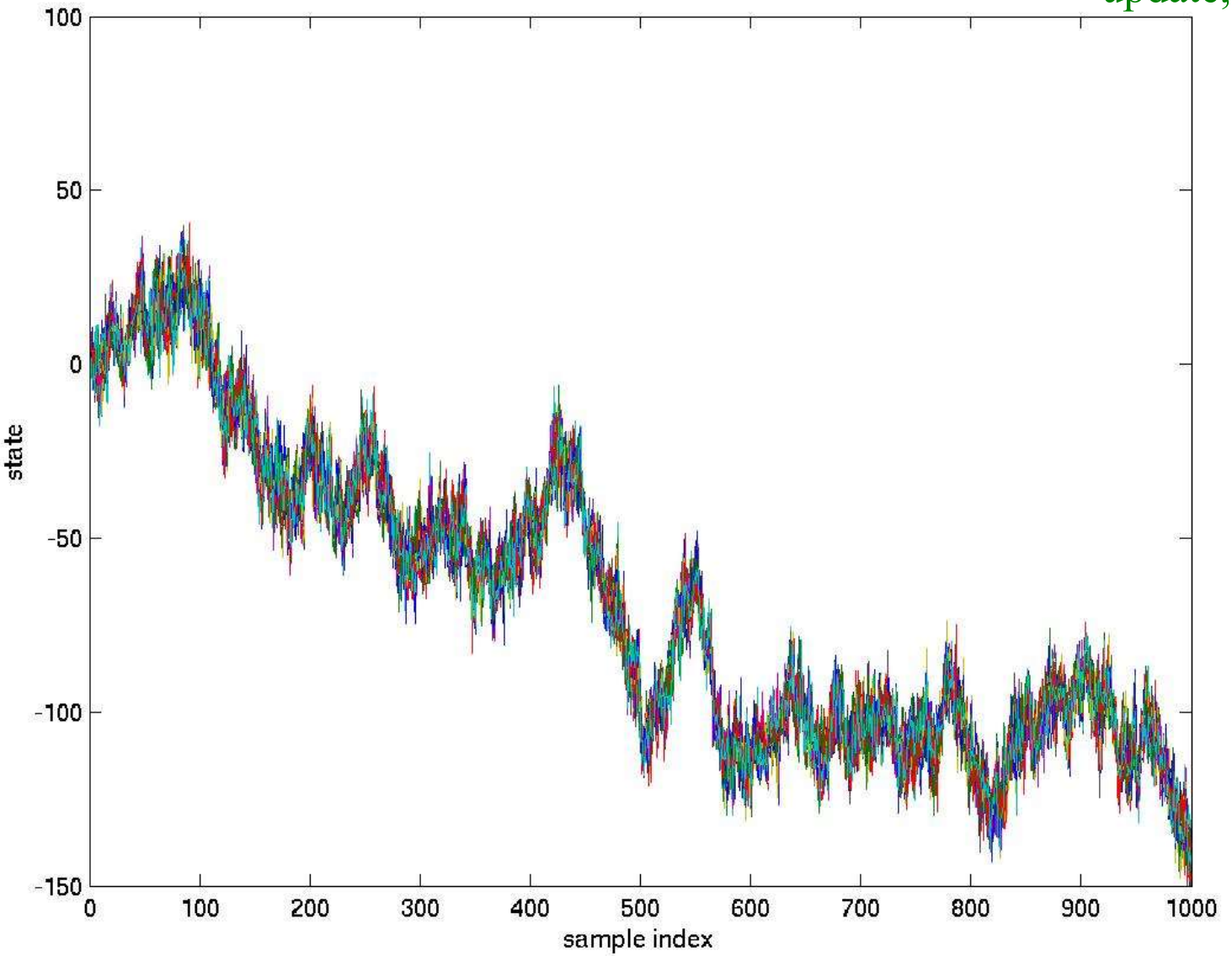


Updated Probs Until T = 1000



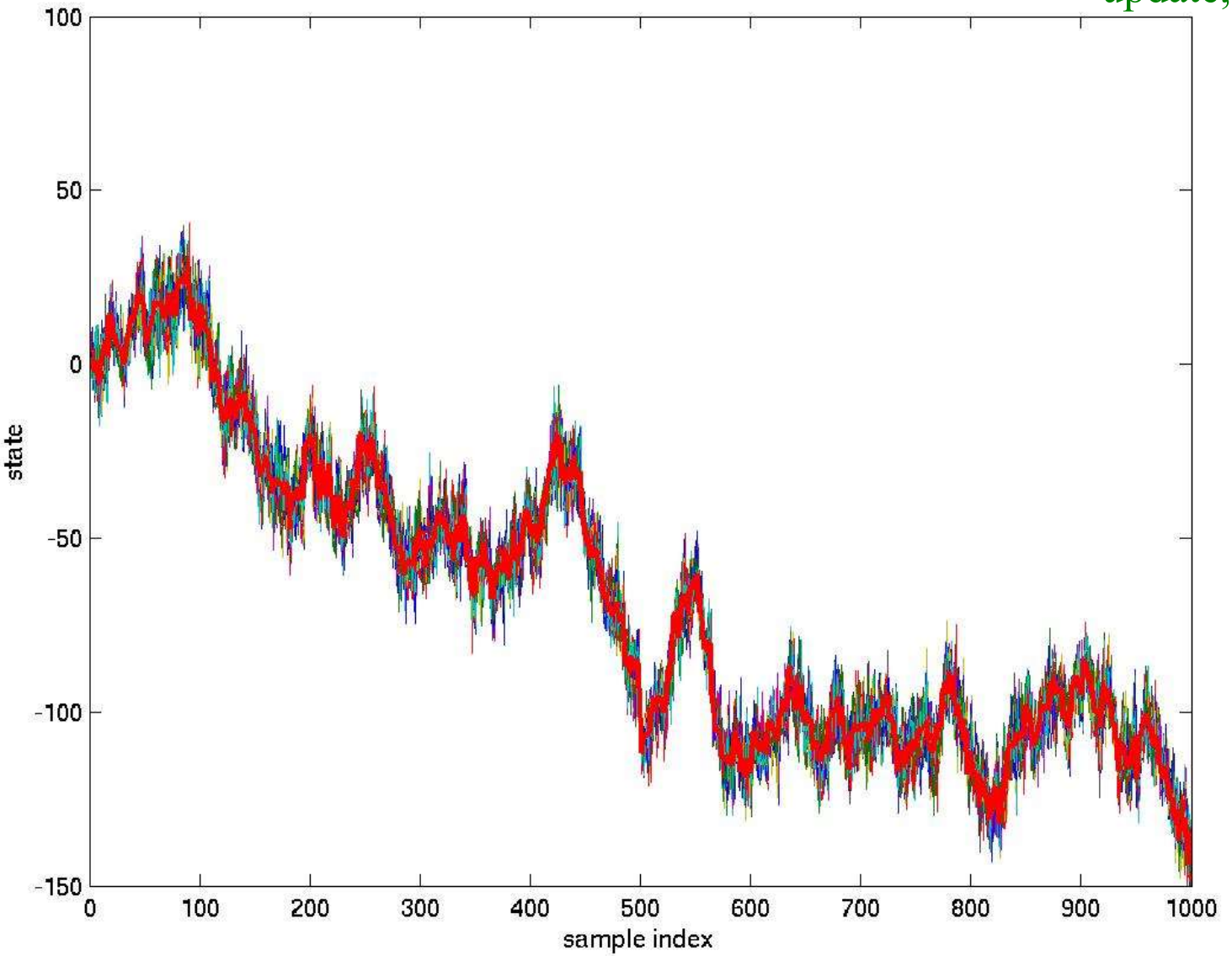
Updated Probs Until T = 1000

update, t <= 1000

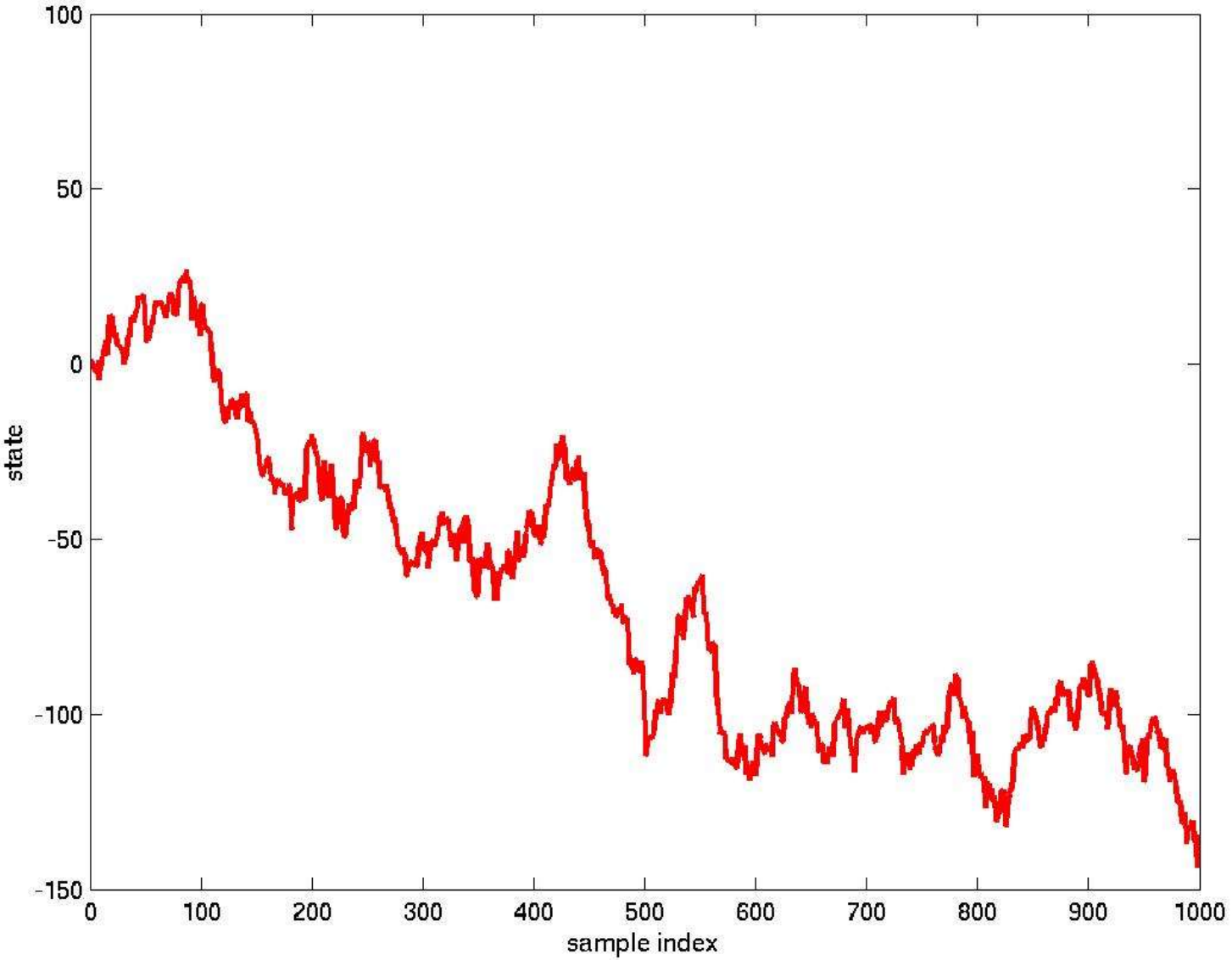


Updated Probs: Top View

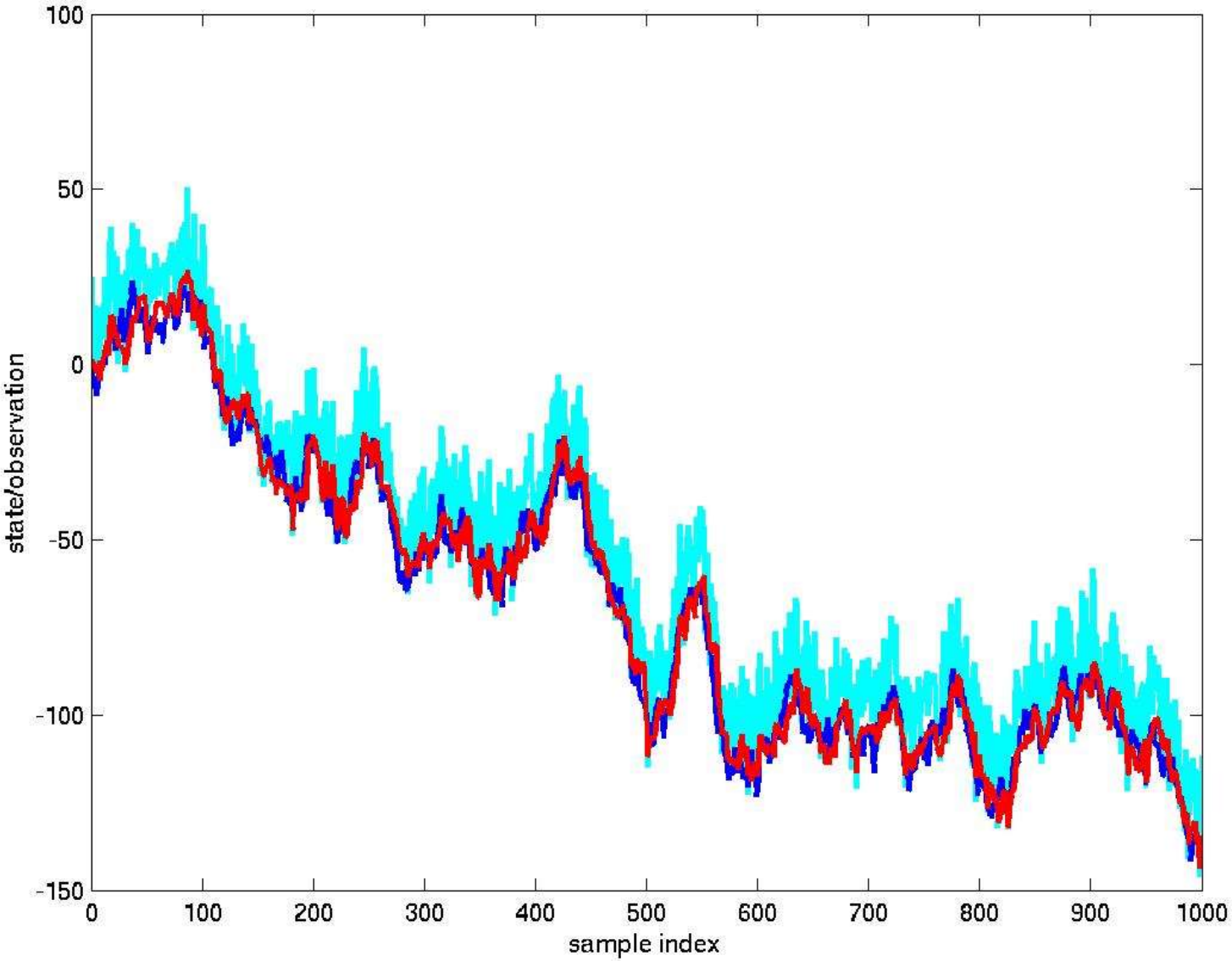
update, t <= 1000



ESTIMATED STATE



Observation, True States, Estimate



Particle Filtering

- Generally quite effective in scenarios where EKF/UKF may not be applicable
  - Potential applications include tracking and edge detection in images!
  - Not very commonly used however
- Highly dependent on sampling
  - A large number of samples required for accurate representation
  - Samples may not represent mode of distribution
  - Some distributions are not amenable to sampling
    - Use importance sampling instead: Sample a Gaussian and assign non-uniform weights to samples

Prediction filters

- HMMs
- Continuous state systems
  - Linear Gaussian: Kalman
  - Nonlinear Gaussian: Extended Kalman
  - Non-Gaussian: Particle filtering

- EKFs are the most commonly used kalman filters..