Machine Learning for Signal Processing

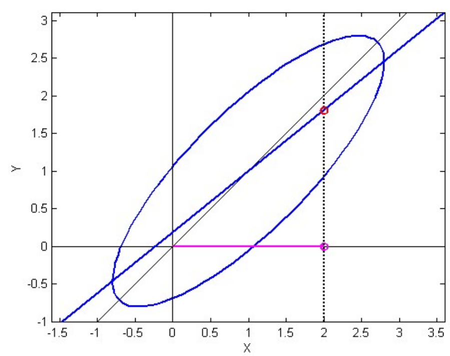
Predicting and Estimation from Time Series

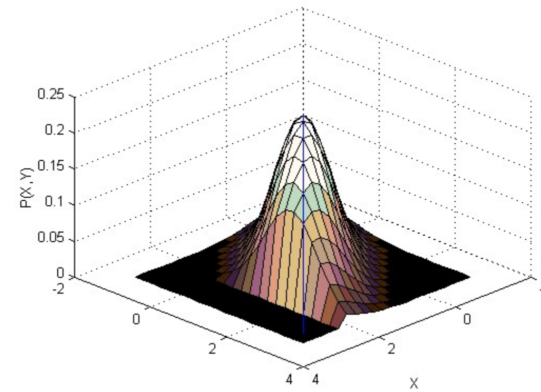
Bhiksha Raj

Preliminaries : P(y|x) for Gaussian

- If P(x,y) is Gaussian:

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





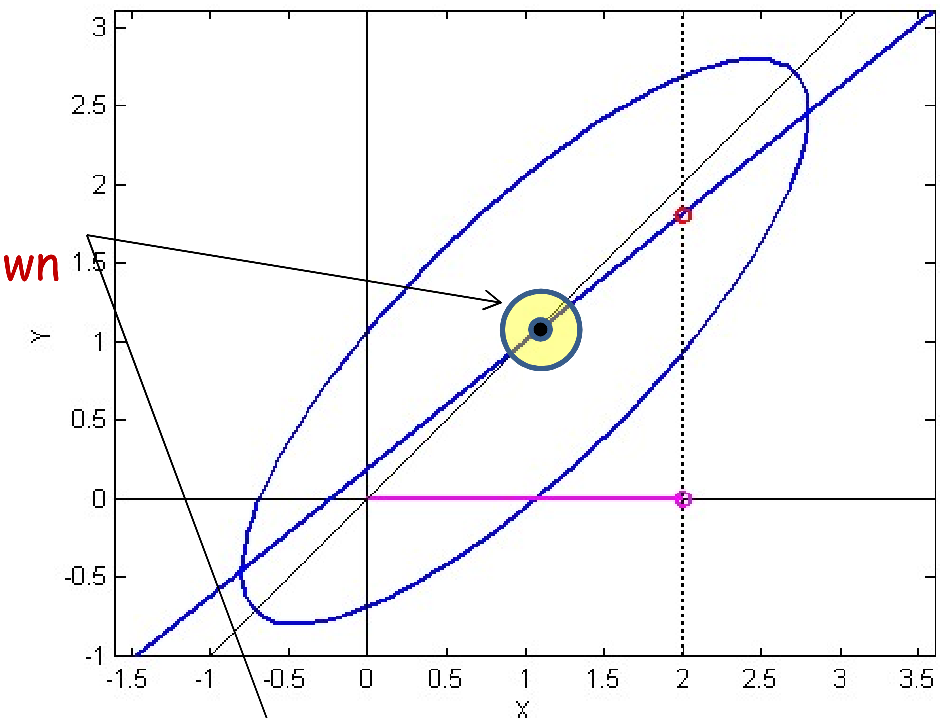
- The conditional probability of y given x is also Gaussian
  - The slice in the figure is Gaussian

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

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

Preliminaries : P(y|x) for Gaussian

Best guess for Y
when X is not known



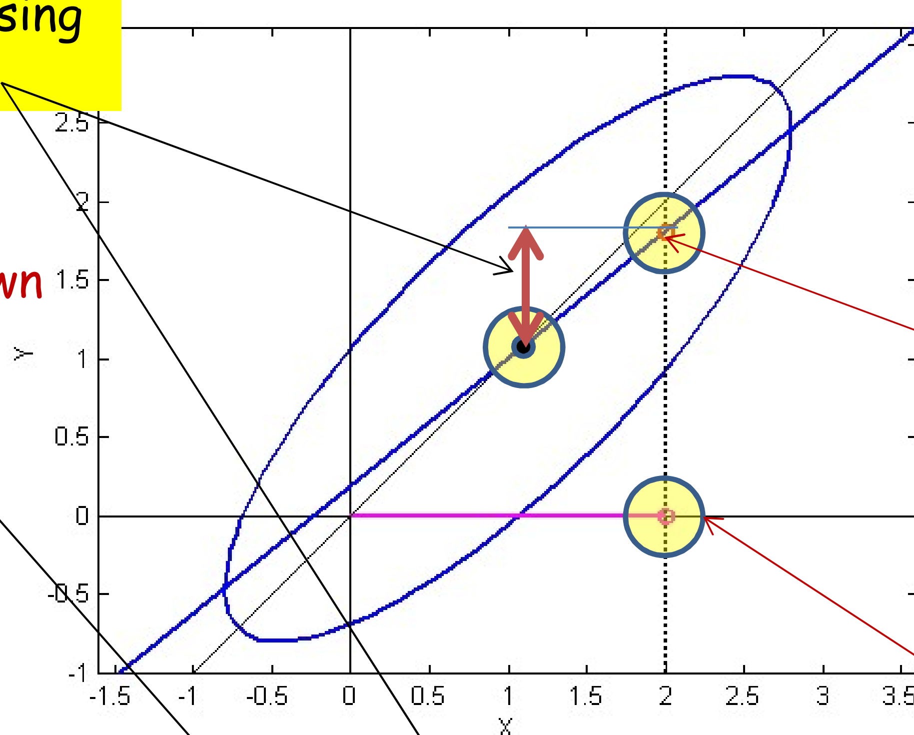
$$P(y \mid x) = N(\mu_y + C_{yx}C_{xx}^{-1}(x - \mu_x), C_{yy} - C_{yx}C_{xx}^{-1}C_{xy})$$

Preliminaries : P(y|x) for Gaussian

Update guess of Y based on information in X
Correction is 0 if X and Y are uncorrelated, i.e $C_{yx} = 0$

Correction of Y using information in X

Best guess for Y when X is not known



Mean of Y given X

Given X value

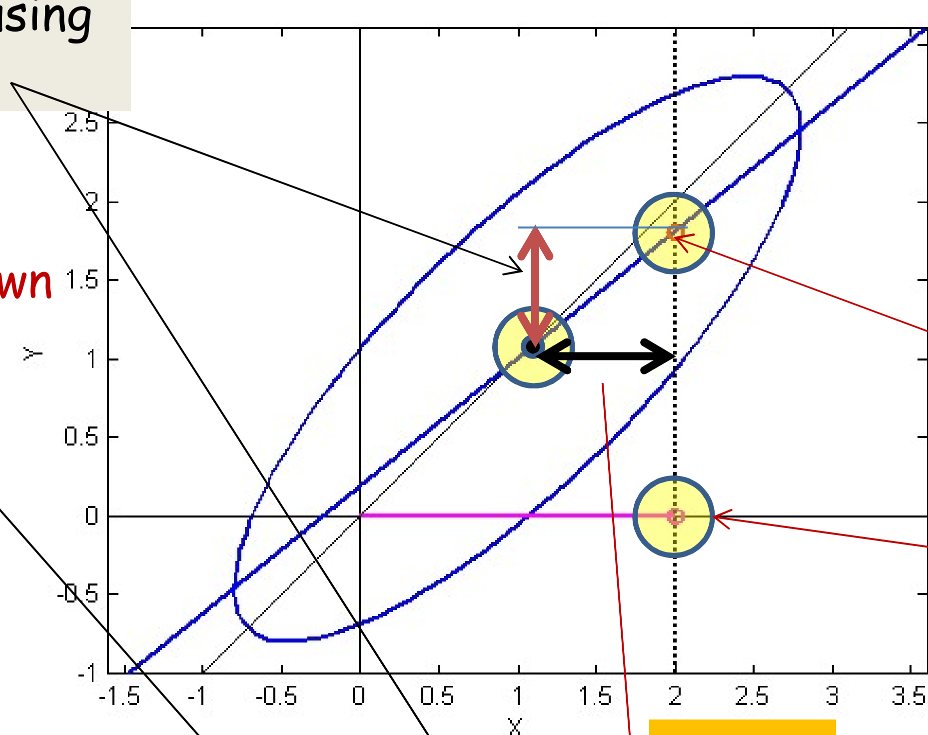
$$P(y \mid x) = N(\mu_y + C_{yx} C_{xx}^{-1} (x - \mu_x), C_{yy} - C_{yx} C_{xx}^{-1} C_{xy})$$

Preliminaries : P(y|x) for Gaussian

Correction to Y = slope * (offset of X from mean)

Correction of Y using information in X

Best guess for Y when X is not known



Mean of Y given X

Given X value

offset

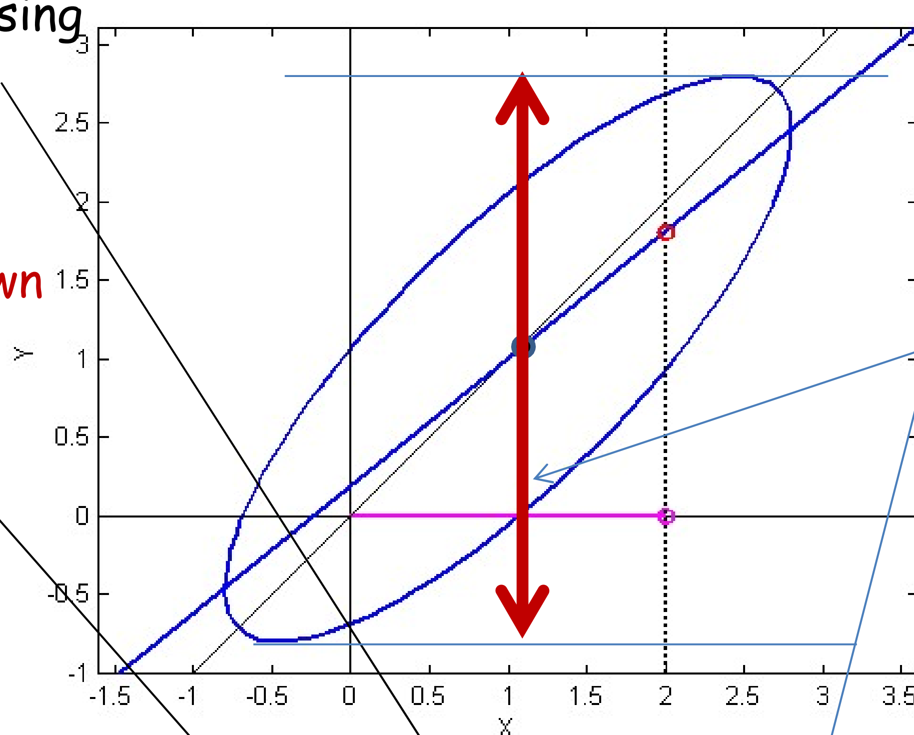
$$P(y|x) = N(\mu_y + C_{yx}C_{xx}^{-1}(x-\mu_x), C_{yy} - C_{yx}C_{xx}^{-1}C_{xy})$$

Slope

Preliminaries : P(y|x) for Gaussian

Correction of Y using information in X

Best guess for Y when X is not known



Uncertainty in Y when X is not known

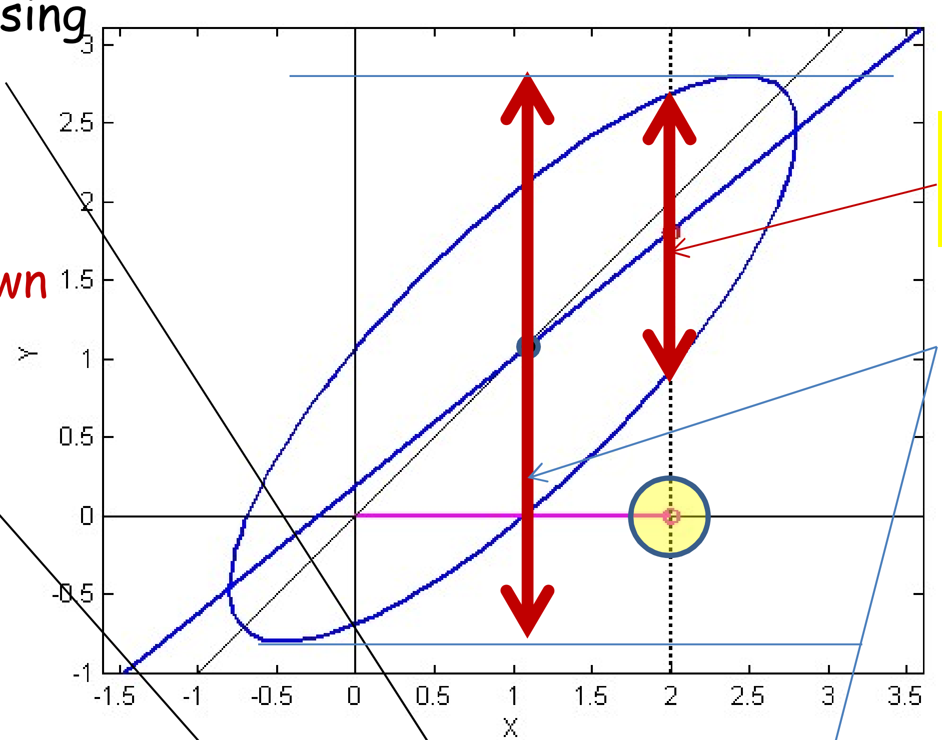
$$P(y \mid x) = N(\mu_y + C_{yx} C_{xx}^{-1} (x - \mu_x), C_{yy} - C_{yx} C_{xx}^{-1} C_{xy})$$

Preliminaries : P(y|x) for Gaussian

Shrinkage of variance is 0 if X and Y are uncorrelated, i.e $C_{yx} = 0$

Correction of Y using information in X

Best guess for Y when X is not known



Reduced uncertainty from knowing X

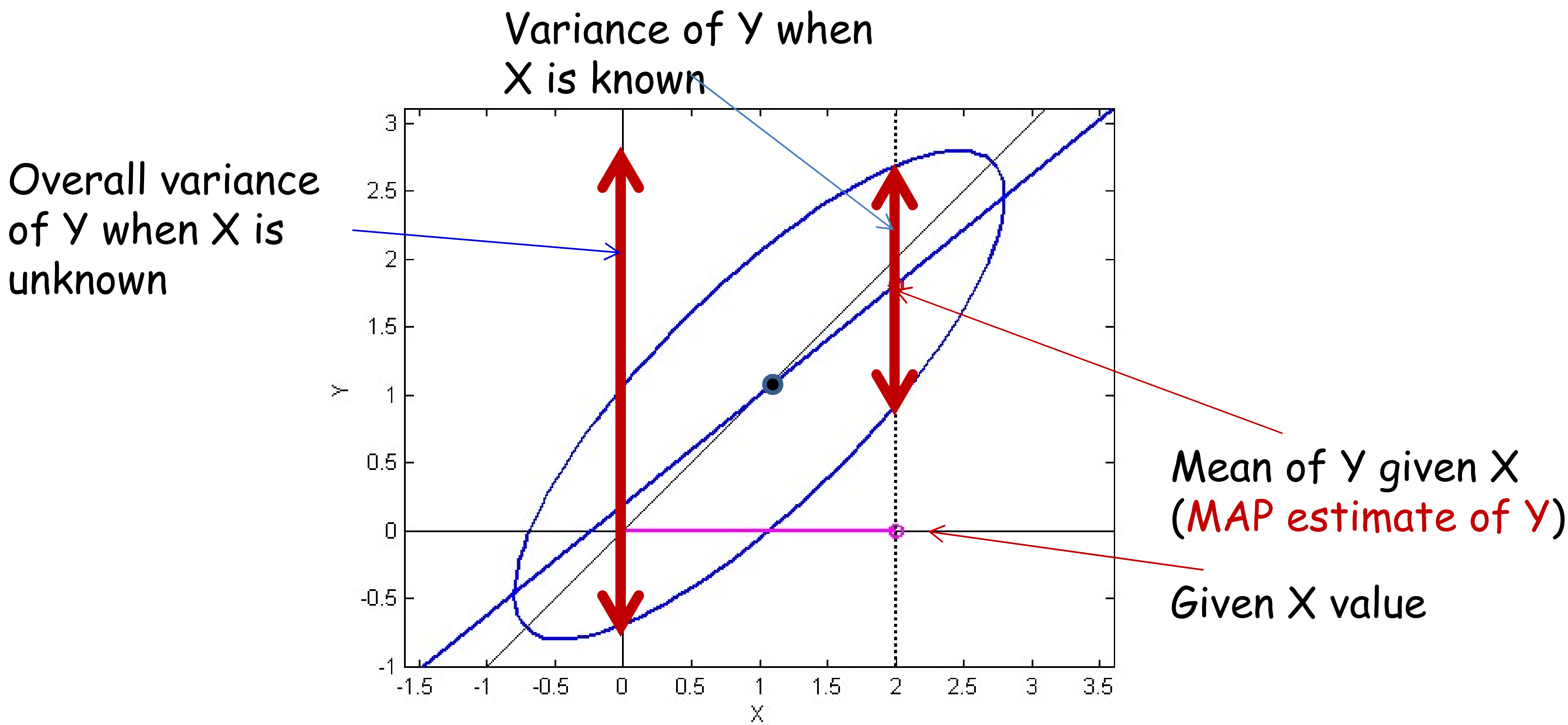
Uncertainty in Y when X is not known

Shrinkage of uncertainty from knowing X

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

Preliminaries : P(y|x) for Gaussian

Knowing X modifies the mean of Y and shrinks its variance



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

Background: Sum of Gaussian RVs

$$O = AS + \varepsilon$$

$$S \sim N(\mu_s, \Theta_s) \qquad \varepsilon \sim N(\mu_\varepsilon, \Theta_\varepsilon)$$

- Consider a random variable *O* obtained as above

- The expected value of *O* is given by

$$E[O] = E[AS + \varepsilon\,] = A\mu_s + \mu_\varepsilon$$

- Notation:

$$E[O] = \mu_O$$

Background: Sum of Gaussian RVs

$$O = AS + \varepsilon$$

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

- The variance of O is given by

$$Var(O) = \Theta_O = E[(O - \mu_O)(O - \mu_O)^T]$$

- This is just the sum of the variance of $\boldsymbol{AS}$ and the variance of $\boldsymbol{\varepsilon}$

$$\Theta_O = A\Theta_S A^{\mathrm{T}} + \Theta_\varepsilon$$

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*:

$$P(O) = N(A\mu_s + \mu_\varepsilon, A\Theta_S A^T + \Theta_\varepsilon)$$

Background: Sum of Gaussian RVs

$$O = AS + \varepsilon$$

$$S \sim N(\mu_s, \Theta_s) \qquad \varepsilon \sim N(\mu_\varepsilon, \Theta_\varepsilon)$$

- The *cross-correlation* between *O* and *S*

$$\Theta_{OS} = E[(O - \mu_O)(S - \mu_s)^T]$$
$$= E[(A(S - \mu_s) + (\varepsilon - \mu_\varepsilon))(S - \mu_s)^T]$$
$$= E[A(S - \mu_s)(S - \mu_s)^T + (\varepsilon - \mu_\varepsilon)(S - \mu_s)^T]$$
$$= AE[(S - \mu_s)(S - \mu_s)^T] + E[(\varepsilon - \mu_\varepsilon)(S - \mu_s)^T]$$
$$= AE[(S - \mu_s)(S - \mu_s)^T]$$

- $= A\,\Theta_s$

- The cross-correlation between *O* and *S* is

$$\Theta_{OS} = A\Theta_S$$
$$\Theta_{SO} = \Theta_S A^T$$

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

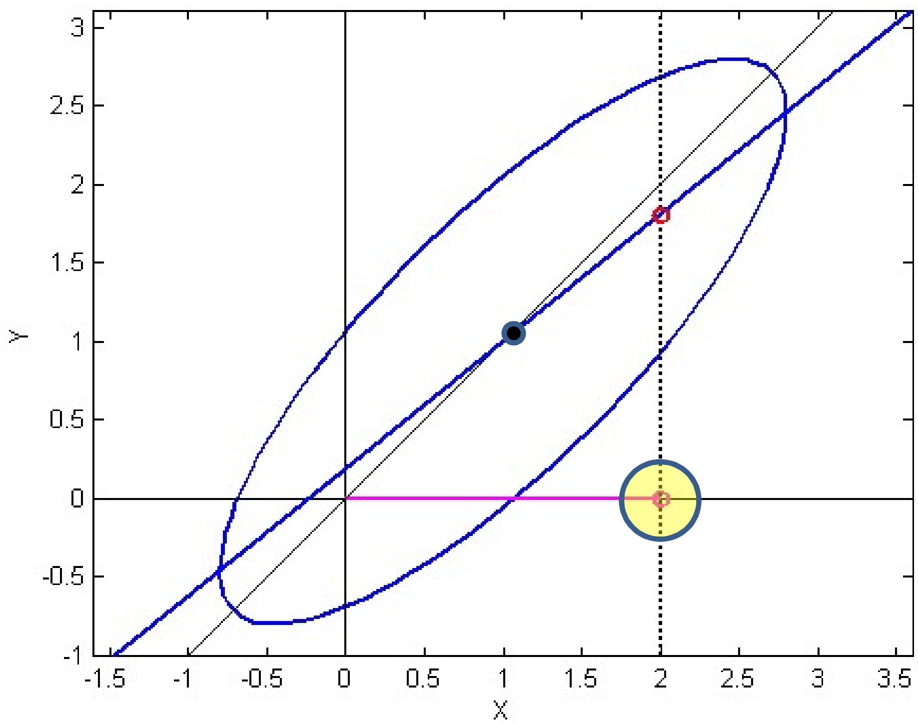
$$P(Z) = P(O, S) = N(\mu_Z, \Theta_Z)$$

- 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)



$$O = AS + \varepsilon$$

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

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

The little parable

You've been kidnapped

And blindfolded

You can only *hear* the car
You must find your way back home from wherever they drop you off

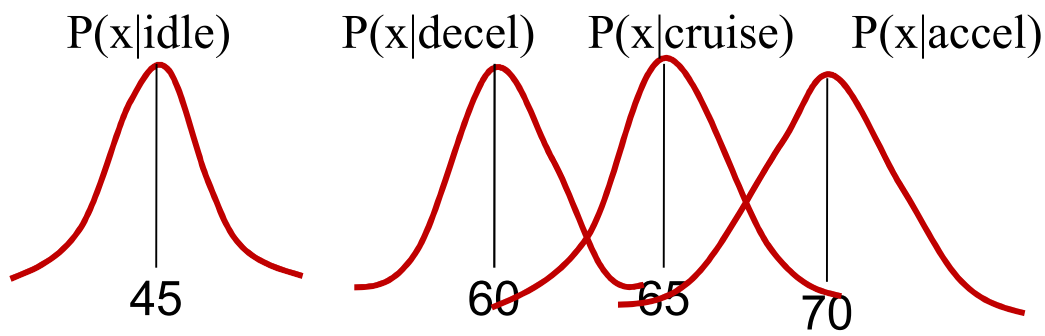
Kidnapped!

- Determine by only *listening* to a running automobile, if it is:
  - Idling; or
  - Travelling at constant velocity; or
  - Accelerating; or
  - Decelerating
- You only record energy level (SPL) in the sound
  - The SPL is measured once per second

What we know

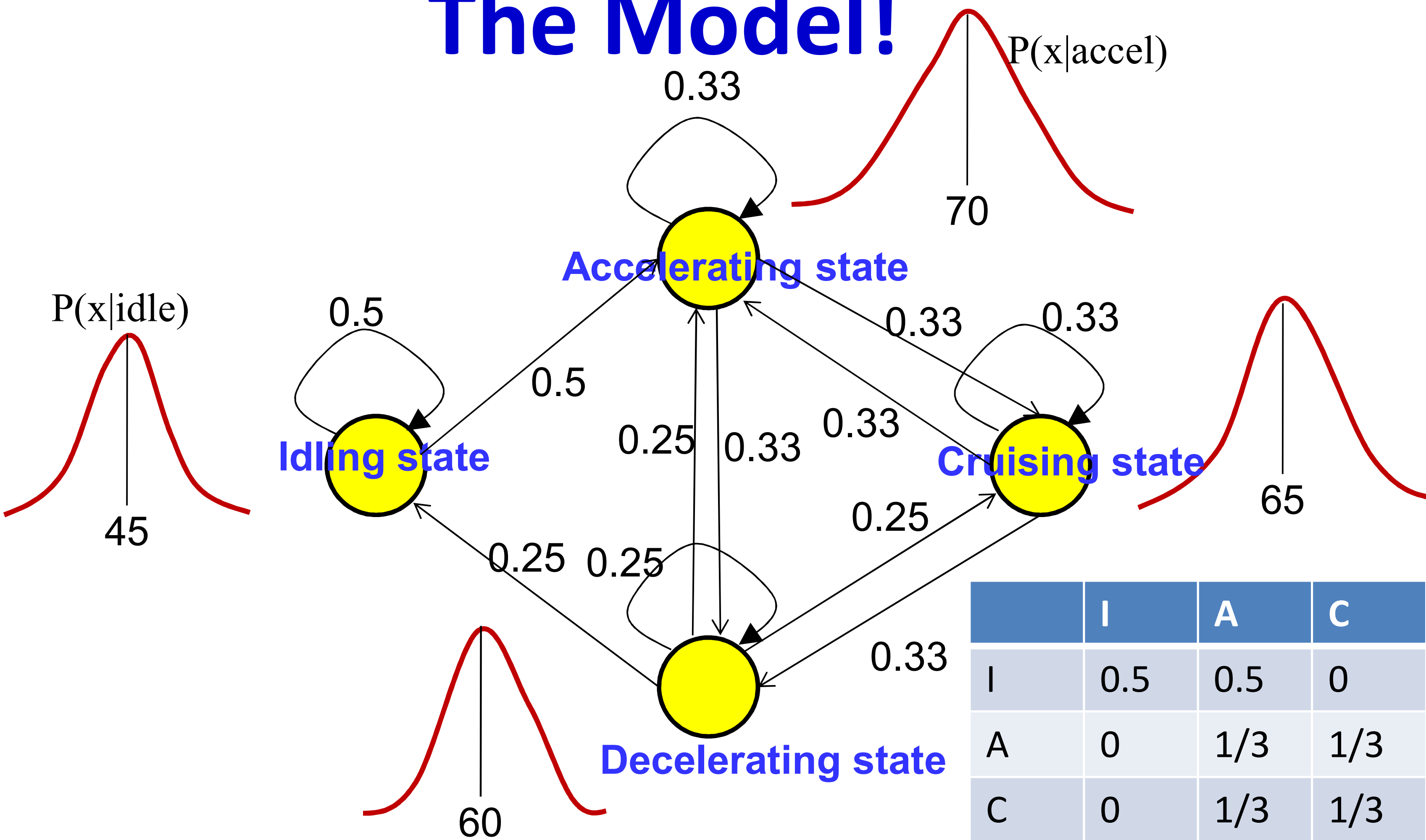
- An automobile that is at rest can accelerate, or continue to stay at rest
- An accelerating automobile can hit a steady-state velocity, continue to accelerate, or decelerate
- A decelerating automobile can continue to decelerate, come to rest, cruise, or accelerate
- A automobile at a steady-state velocity can stay in steady state, accelerate or decelerate

What else we know



- The probability distribution of the SPL of the sound is different in the various conditions
  – As shown in figure
    - In reality, depends on the car
- The distributions for the different conditions overlap
  – Simply knowing the current sound level is not enough to know the state of the car

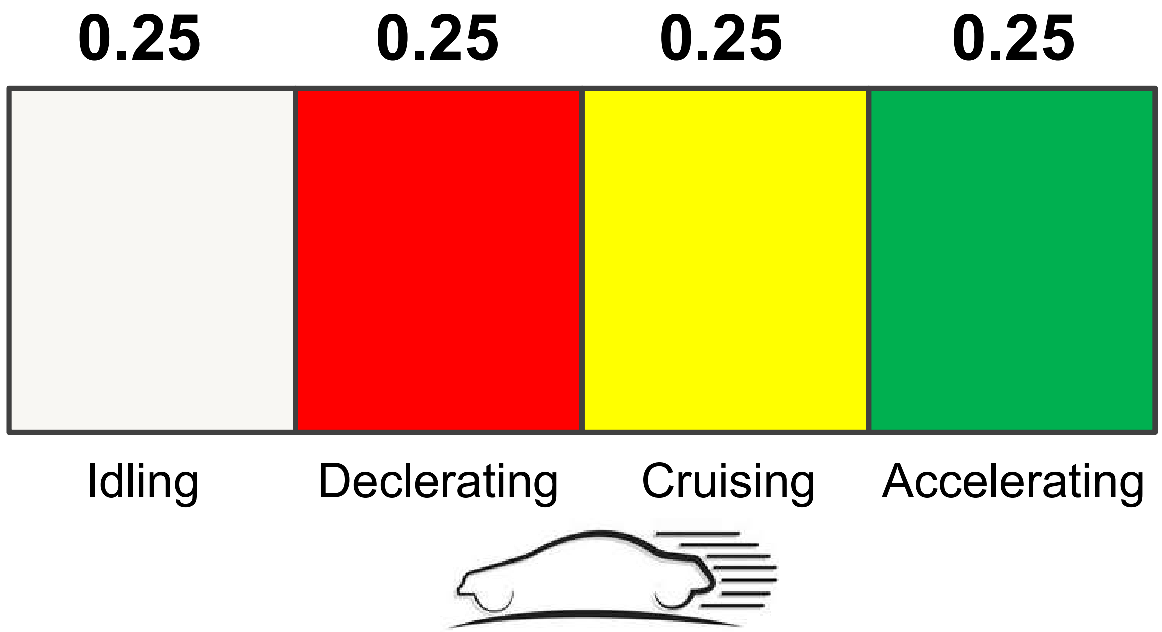
The Model!



	I	A	C	D
I	0.5	0.5	0	0
A	0	1/3	1/3	1/3
C	0	1/3	1/3	1/3
D	0.25	0.25	0.25	0.25

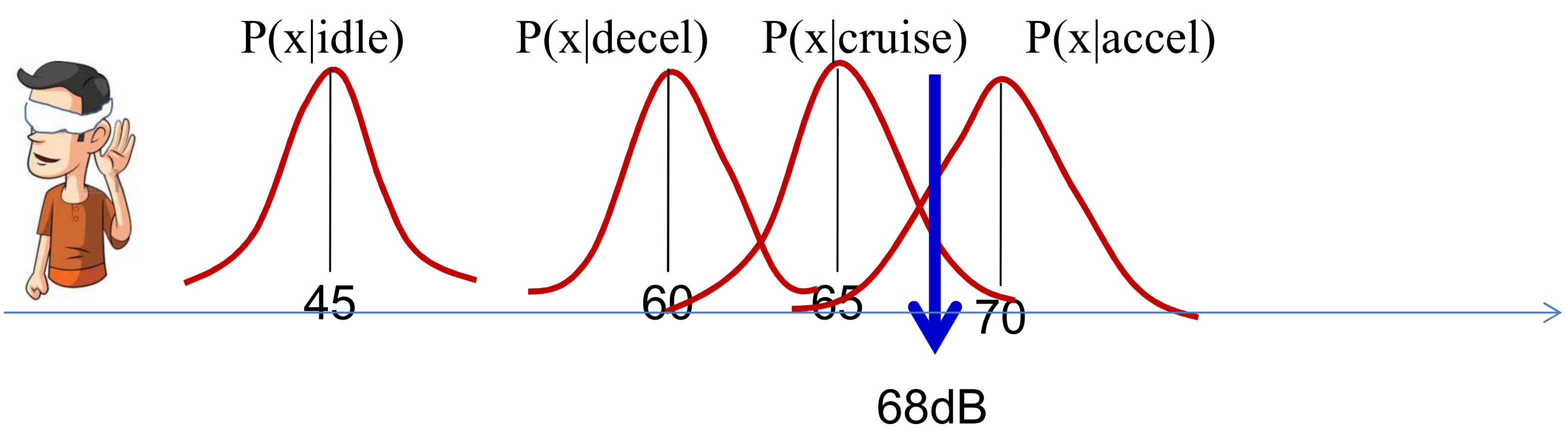
- The state-space model
  - Assuming all transitions from a state are equally probable
  - This is a Hidden Markov Model!

Estimating the state at T = 0-



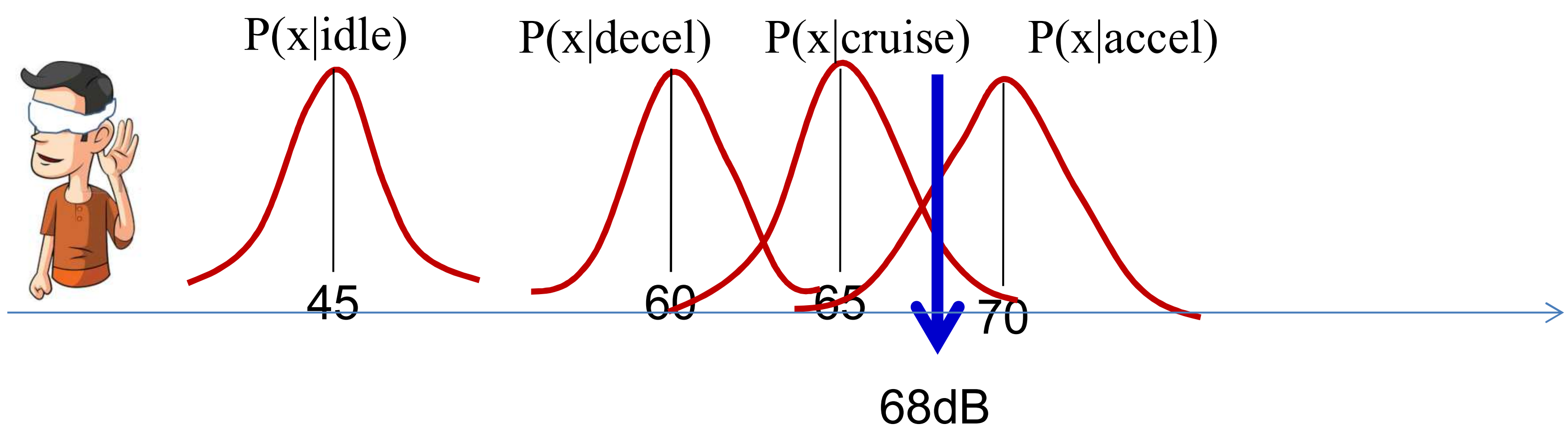
- A T=0, before the first observation, we know nothing of the state
  - Assume all states are equally likely

The first observation: T=0



- At T=0 you observe the sound level x_0 = 68dB SPL
  - The observation modifies our belief in the state of the system

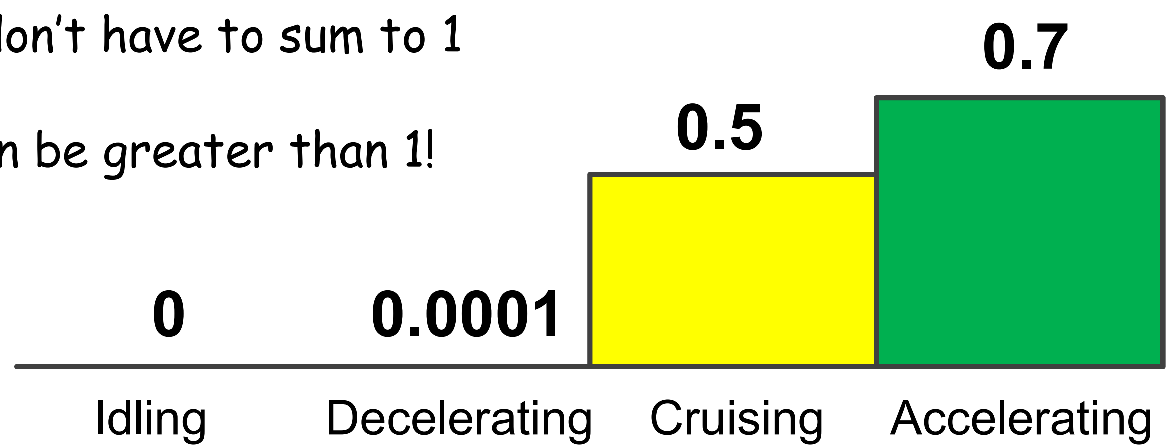
The first observation: T=0



P(x|idle)	P(x|deceleration)	P(x|cruising)	P(x|acceleration)
0	0.0001	0.5	0.7

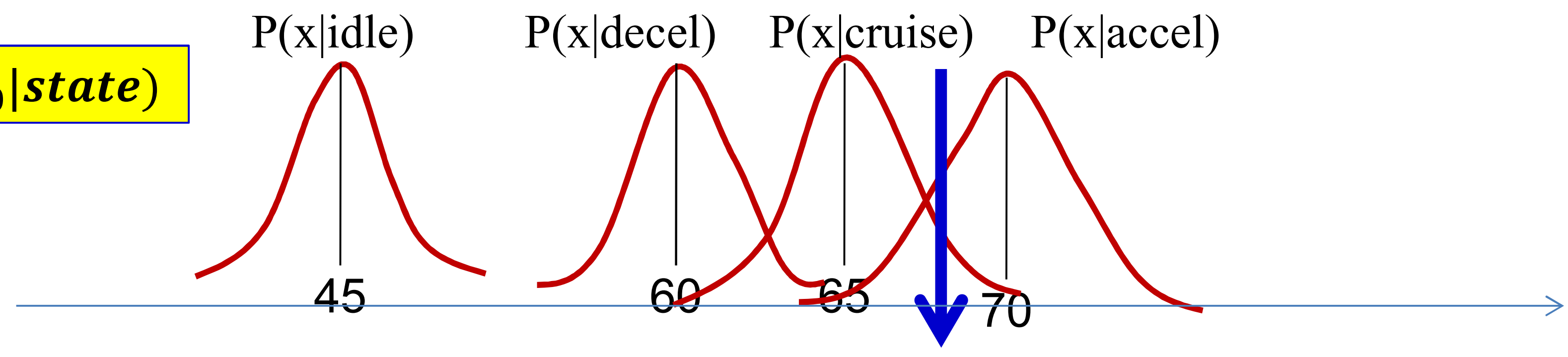
These don't have to sum to 1

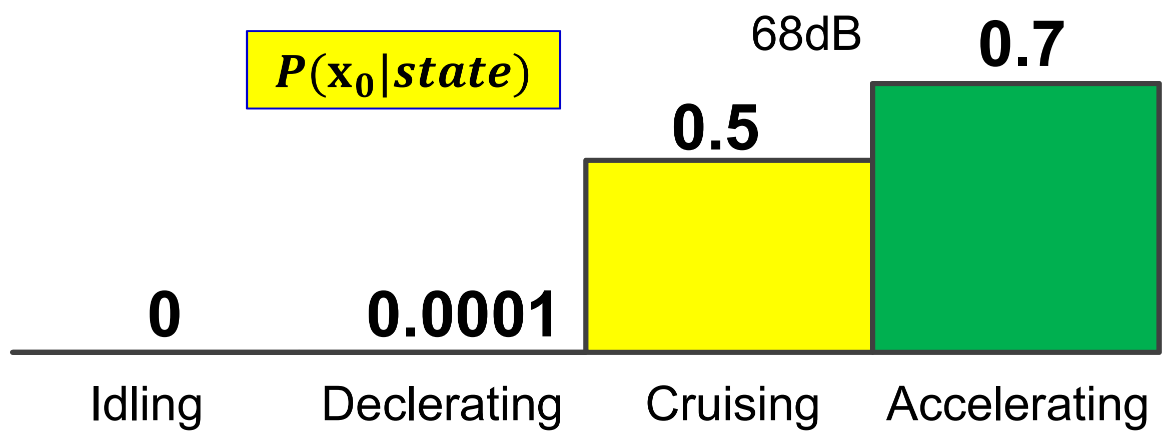
Can even be greater than 1!



The first observation: T=0

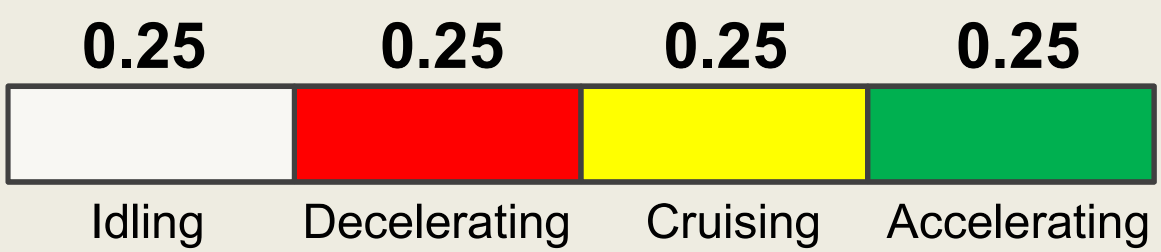
$P(x_0|state)$





$Prior: \quad P(state)$

Remember the prior

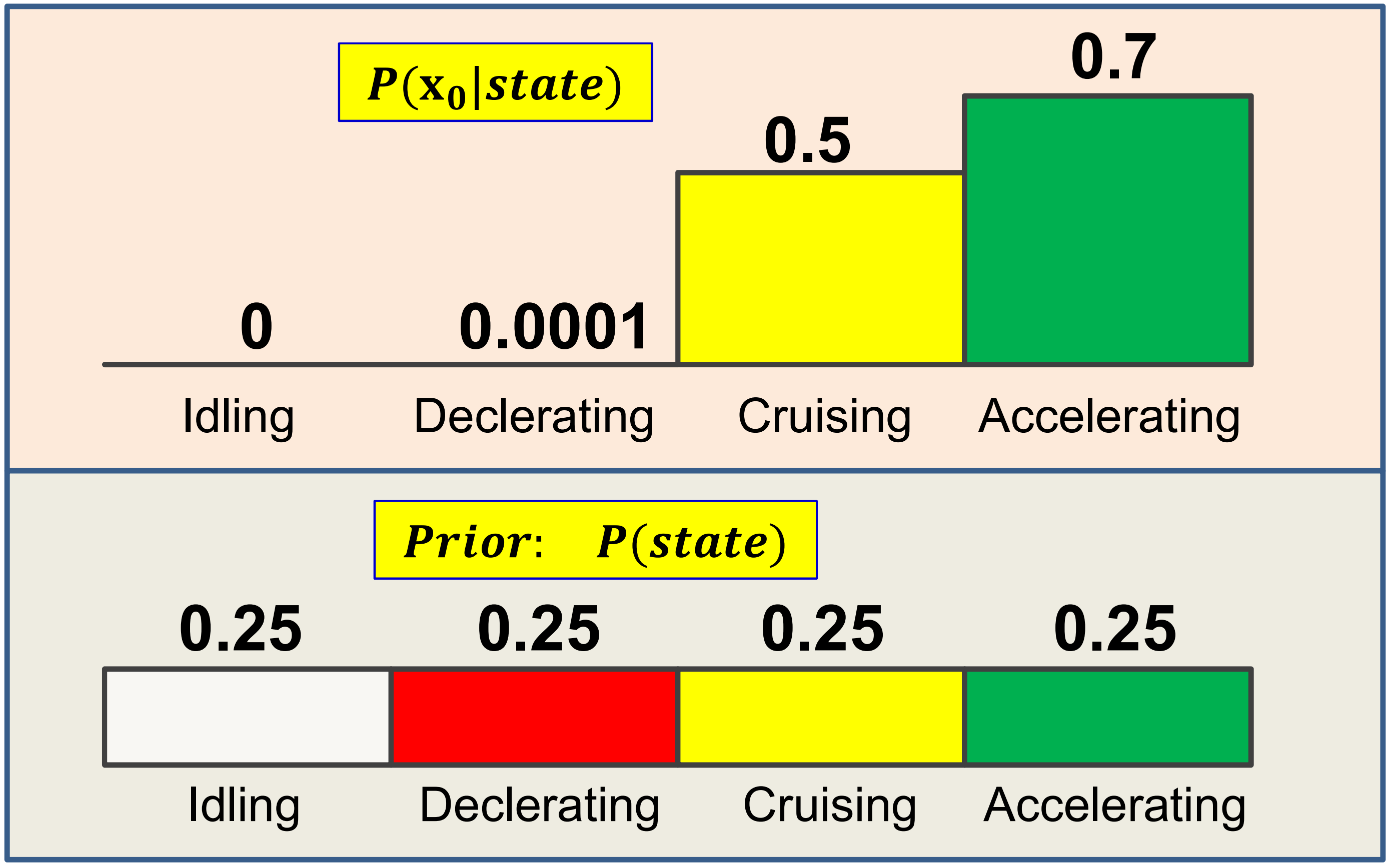


Estimating state *after* at observing $\mathbf{x}_0$

- Combine prior information about state and evidence from observation
- We want $P(state|\mathbf{x}_0)$
- We can compute it using Bayes rule as

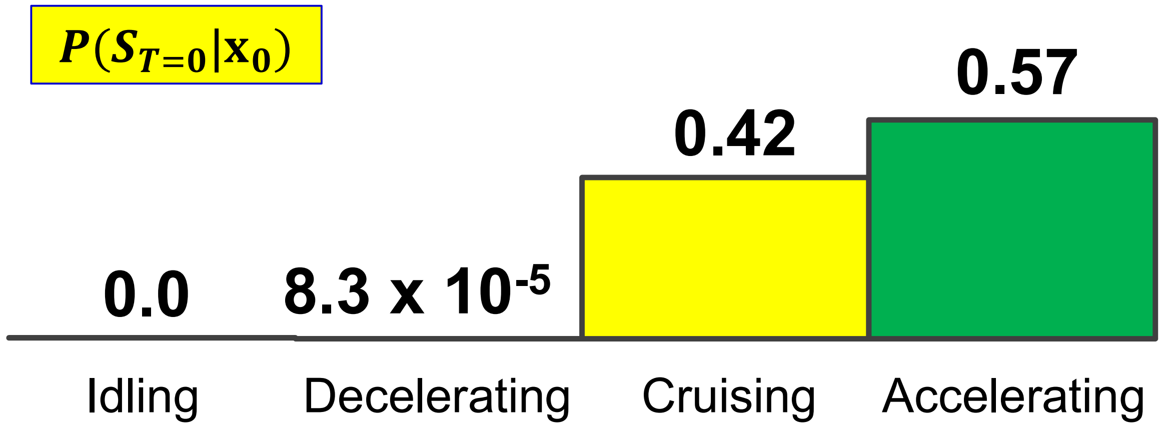
$$P(state|x_0) = \frac{P(state)P(\mathrm{x}_0|state)}{\sum_{state'} P(state')P(\mathrm{x}_0|state')}$$

The Posterior



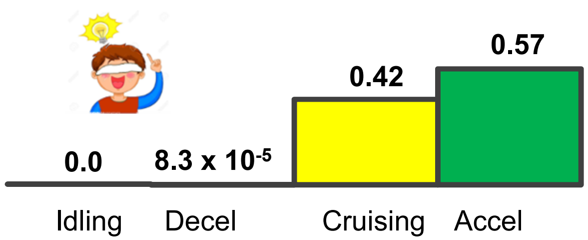
- Multiply the two, term by term, and normalize them so that they sum to 1.0

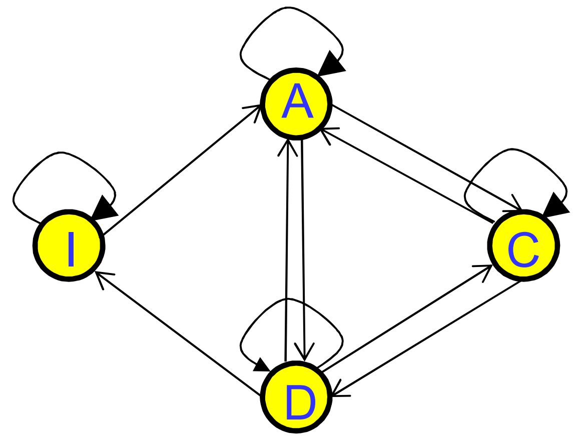
Estimating the state at T = 0+



- At T=0, after the first observation x_0, we update our belief about the states
  - The first observation provided some evidence about the state of the system
  - It modifies our belief in the state of the system

Predicting the state at T=1





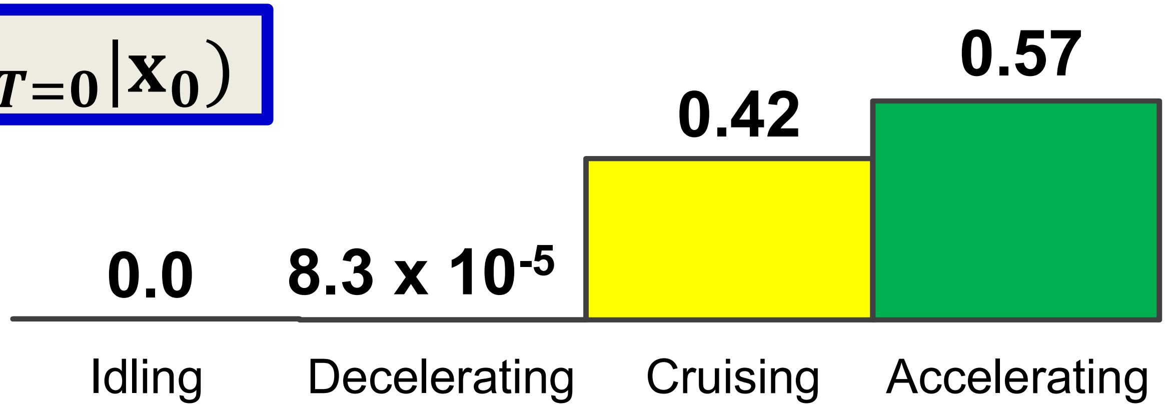
	I	A	C	D
I	0.5	0.5	0	0
A	0	1/3	1/3	1/3
C	0	1/3	1/3	1/3
D	0.25	0.25	0.25	0.25

- Predicting the probability of idling at T=1
  - P(*idling* | *idling*) = 0.5;
  - P(*idling* | *deceleration*) = 0.25
  - P(*idling* at T=1| x_0) =
    $P(I_{T=0}|x_0)\ P(I|I) + P(D_{T=0}|x_0)\ P(I|D) = 2.1 \times 10^{-5}$

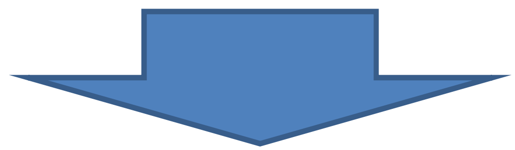
- In general, for any state S
  - $P(S_{T=1}|\mathbf{x}_0) = \sum_{S_{T=0}} P(S_{T=0}|\mathbf{x}_0) P(S_{T=1}|S_{T=0})$

Predicting the state at T = 1

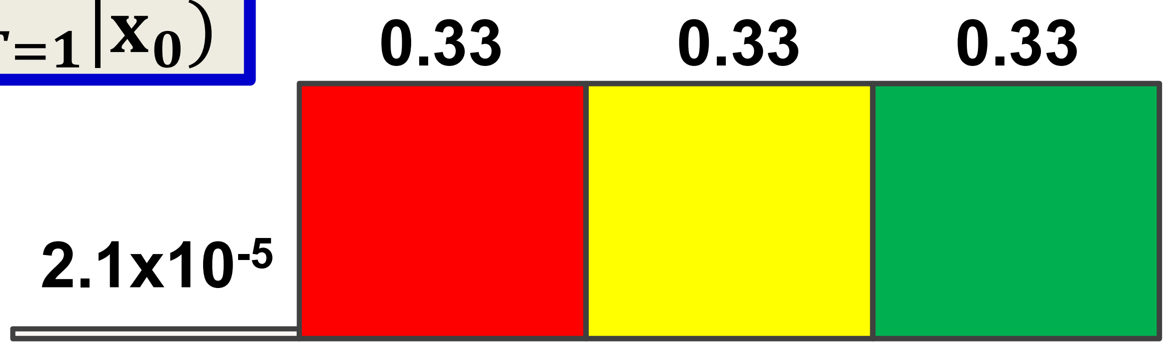
$$P(S_{T=0}|\mathbf{x}_0)$$



$$P(S_{T=1}|\mathbf{x}_0) = \sum_{S_{T=0}} P(S_{T=0}|\mathbf{x}_0)P(S_{T=1}|S_{T=0})$$

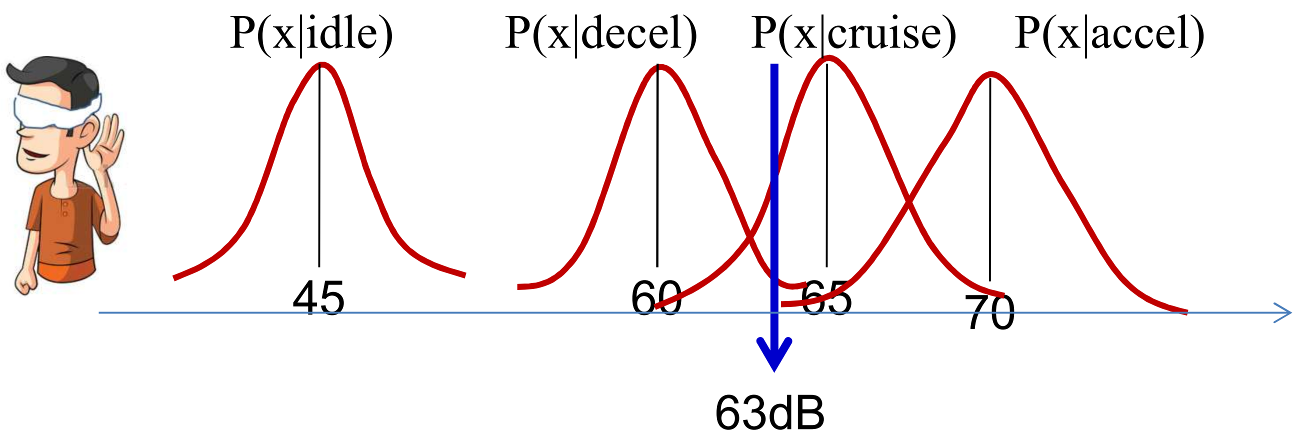


$$P(S_{T=1}|\mathbf{x}_0)$$



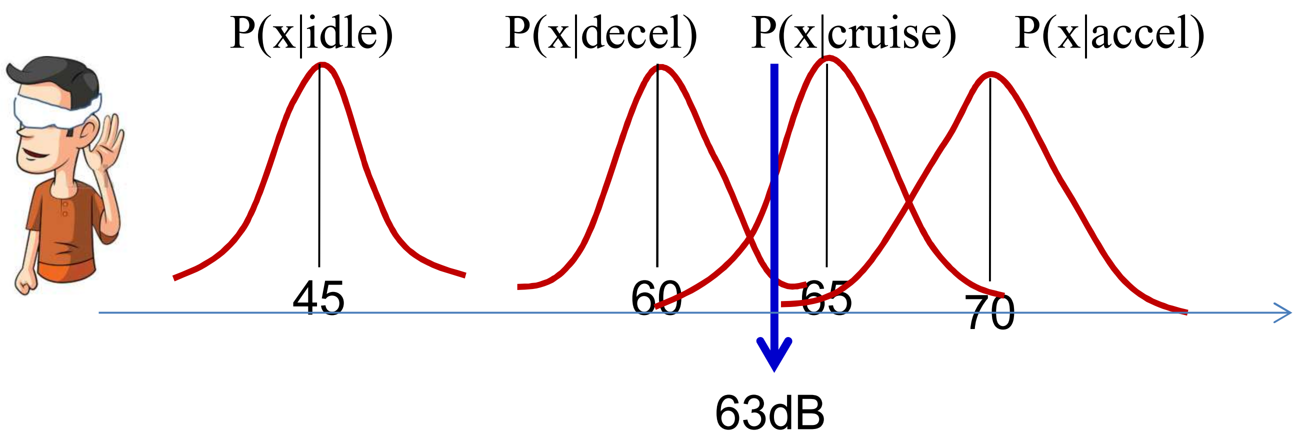
Rounded.
In reality, they sum to 1.0

Updating after the observation at T=1

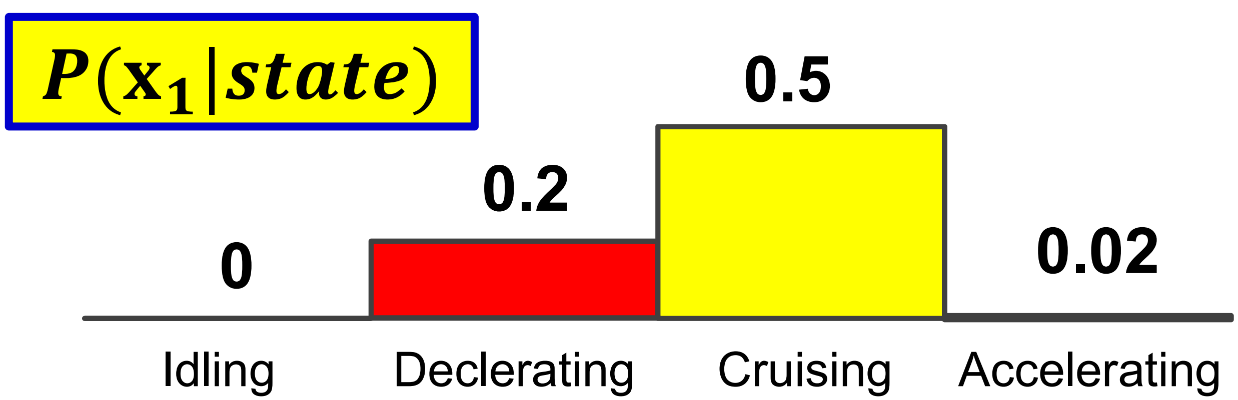


- At T=1 we observe x_1 = 63dB SPL

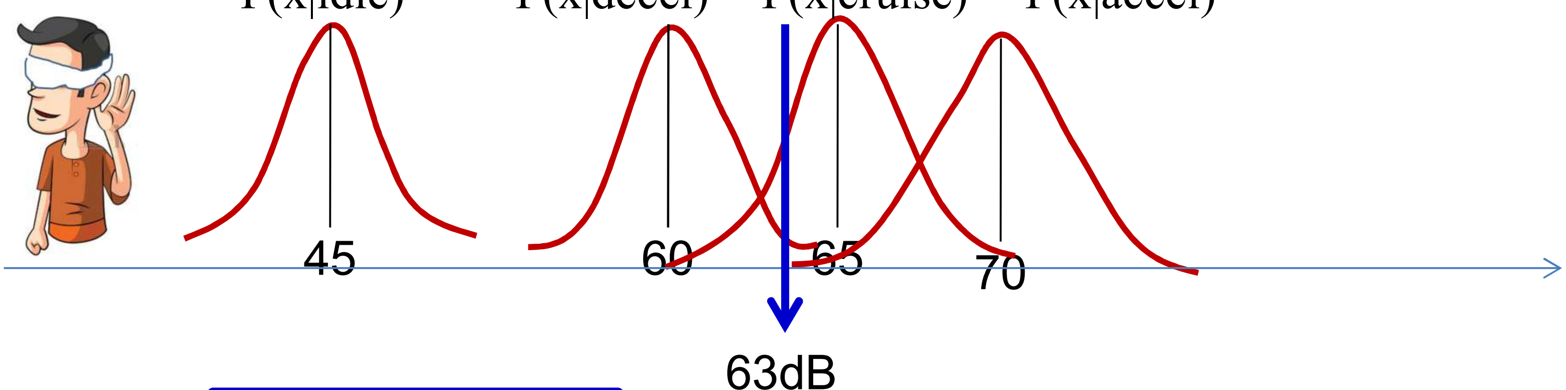
Updating after the observation at T=1

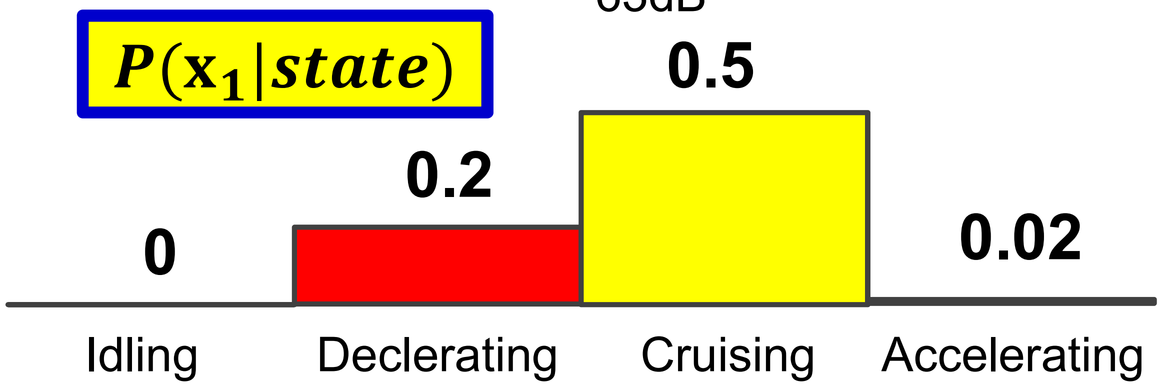


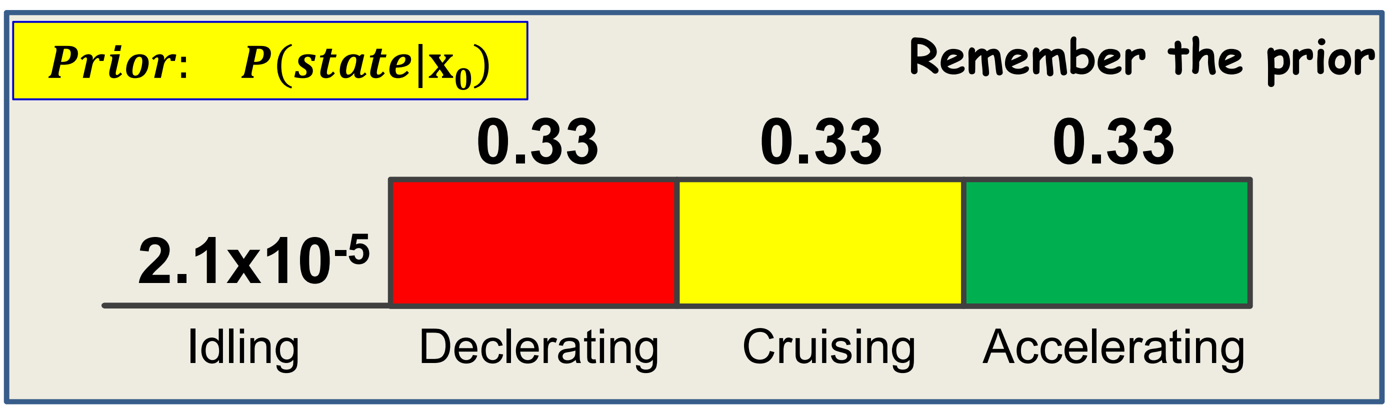
P(x|idle)	P(x|deceleration)	P(x|cruising)	P(x|acceleration)
0	0.2	0.5	0.01



The second observation: T=1







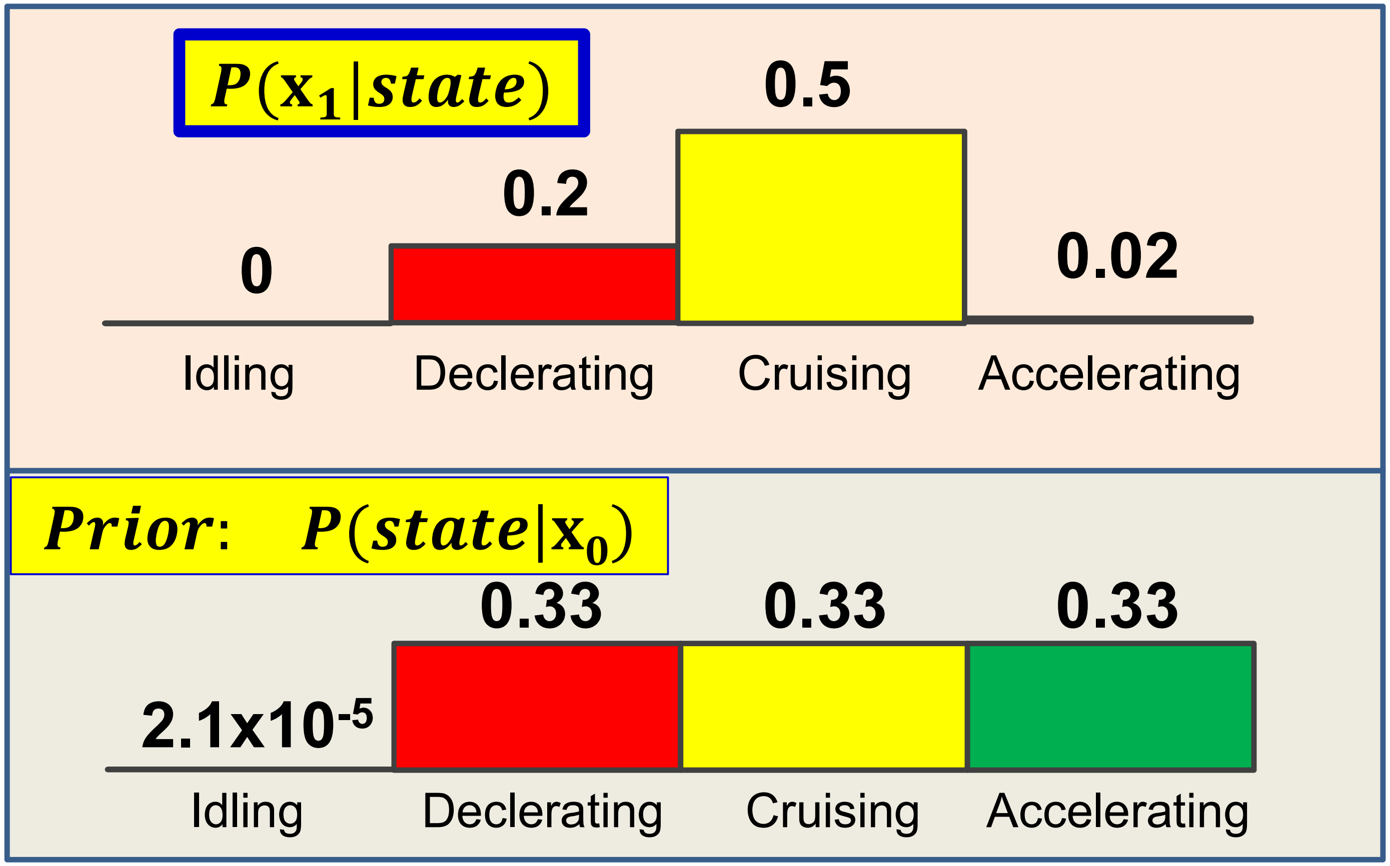
Estimating state *after* at observing $\mathbf{x}_1$

- Combine prior information from the observation at time T=0, AND evidence from observation at T=1 to estimate ***state*** at T=1

- We want $P(state|\mathbf{x}_0, \mathbf{x}_1)$
- We can compute it using Bayes rule as

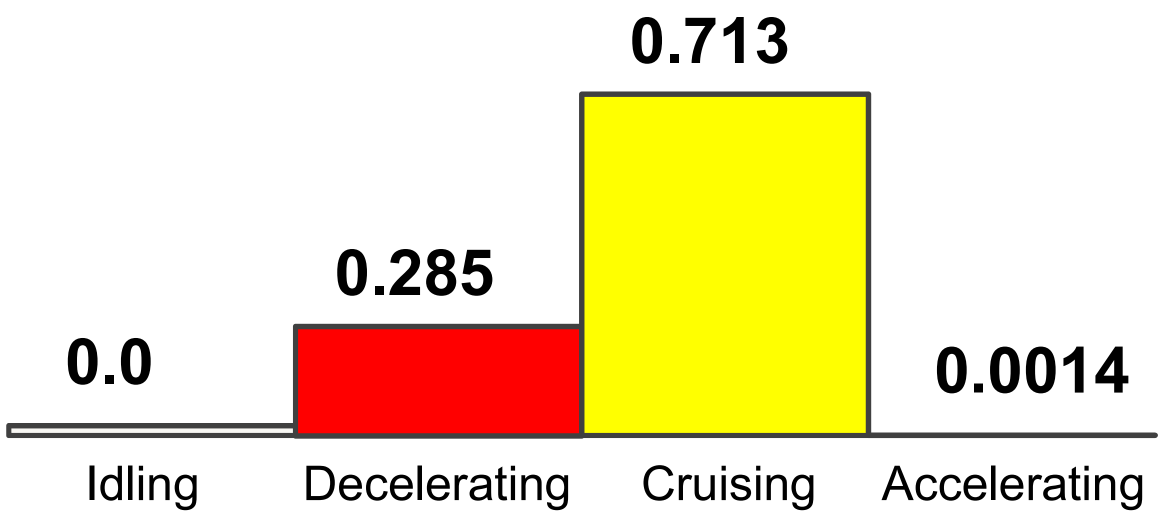
$$P(state|\mathbf{x}_0, \mathbf{x}_1) = \frac{P(state|\mathbf{x}_0)P(\mathbf{x}_1|state)}{\sum_{state'} P(state'|\mathbf{x}_0)P(\mathbf{x}_1|state')}$$

The Posterior at T = 1



- Multiply the two, term by term, and normalize them so that they sum to 1.0

Estimating the state at T = 1+

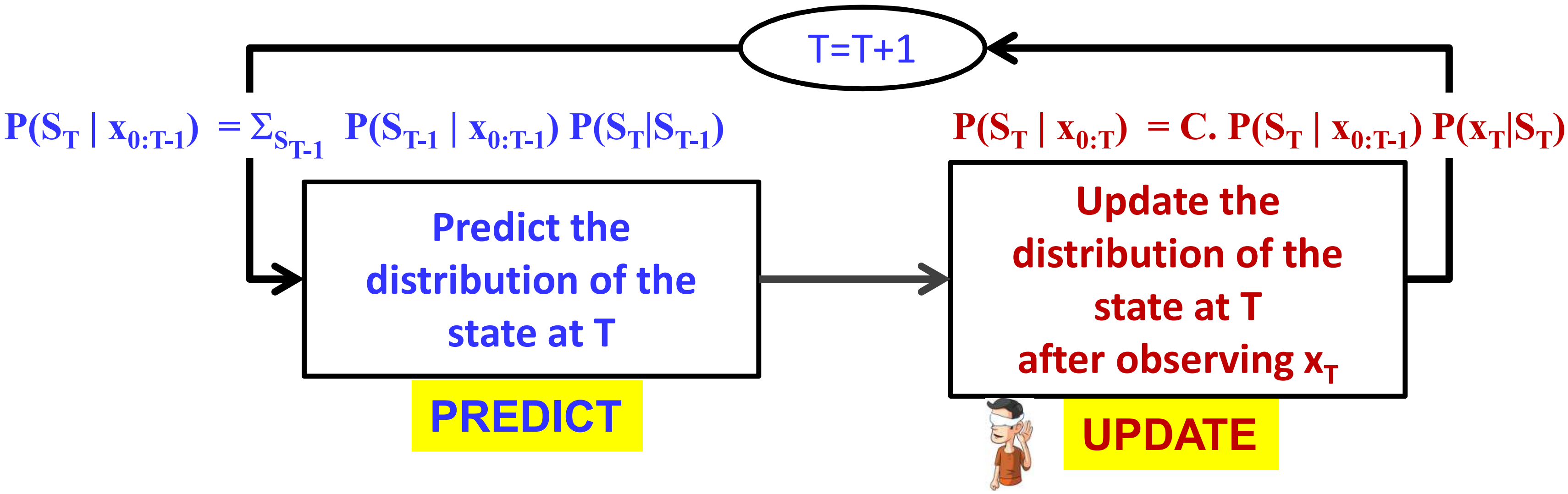


- The updated probability at T=1 incorporates information from both x_0 and x_1
  - It is NOT a local decision based on x_1 alone
  - Because of the Markov nature of the process, the state at T=0 affects the state at T=1
    - x_0 provides evidence for the state at T=1

Overall Process

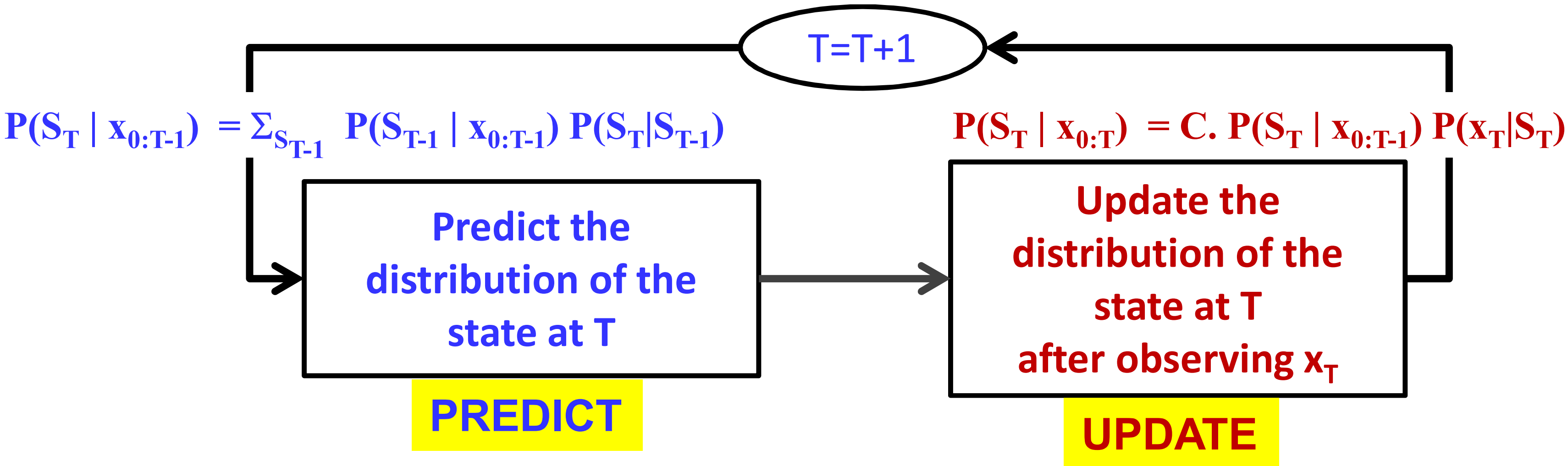
Time	Computation
T=0- : A priori probability	$P(S_0) = P(S)$
T = 0+: Update after X_0	$P(S_0\|X_0) = C.P(S_0)P(X_0\|S_0)$
T=1- (Prediction before X_1)	$P(S_1\|X_0) = \sum_{S_0} P(S_1\|S_0)P(S_0\|X_0)$
T = 1+: Update after X_1	$P(S_1\|X_{0:1}) = C.P(S_1\|X_0)P(X_1\|S_1)$
T=2- (Prediction before X_2)	$P(S_2\|X_{0:1}) = \sum_{S_1} P(S_2\|S_1)P(S_1\|X_{0:1})$
T = 2+: Update after X_2	$P(S_2\|X_{0:2}) = C.P(S_2\|X_{0:1})P(X_2\|S_2)$
…	…
T= t- (Prediction before X_t)	$P(S_t\|X_{0:t-1}) = \sum_{S_{t-1}} P(S_t\|S_{t-1})P(S_{t-1}\|X_{0:t-1})$
T = t+: Update after X_t	$P(S_t\|X_{0:t}) = C.P(S_t\|X_{0:t-1})P(X_t\|S_t)$

Overall procedure



- At T=0 the predicted state distribution is the initial state probability
- At each time T, the current estimate of the distribution over states considers *all* observations $x_0 \ldots x_T$
  - A natural outcome of the Markov nature of the model
- The prediction+update is identical to the forward computation for HMMs to within a normalizing constant

Comparison to Forward Algorithm



- Forward Algorithm:
  - $P(x_{0:T},S_T) = P(x_T|S_T) \Sigma_{S_{T-1}} P(x_{0:T-1}, S_{T-1}) P(S_T|S_{T-1})$

PREDICT

UPDATE

- Normalized:
  - $P(S_T|x_{0:T}) = (\Sigma_{S'_T} P(x_{0:T},S'_T))^{-1} P(x_{0:T},S_T) = C\ P(x_{0:T},S_T)$

Decomposing the Algorithm

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

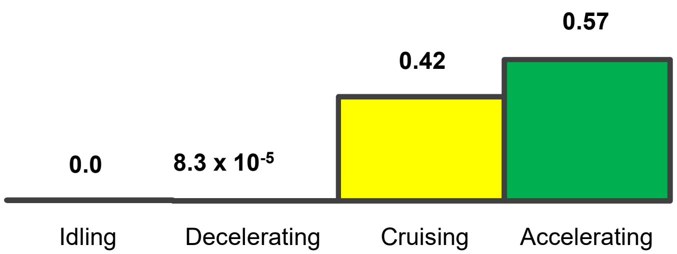
Predict: $P(S_t|X_{0:t-1}) = \sum_{S_{t-1}} P(S_t|S_{t-1}) P(S_{t-1}|X_{0:t-1})$

Update:

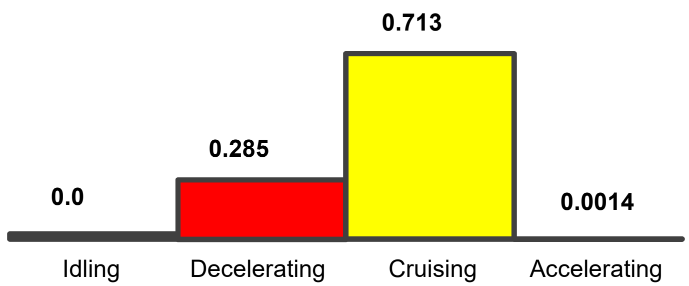
$$P(S_t|X_{0:t}) = \frac{P(S_t|X_{0:t-1}) P(X_t|S_t)}{\sum_S P(S|X_{0:t-1}) P(X_t|S)}$$

Estimating a Unique state

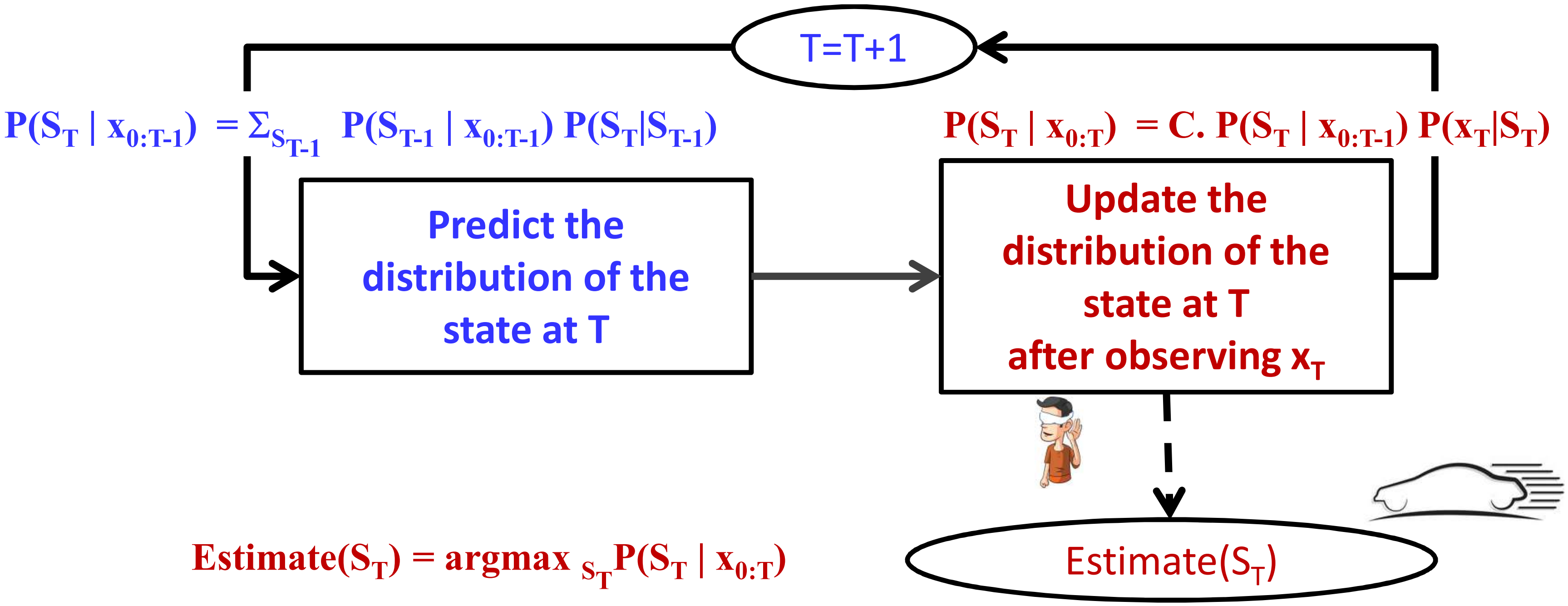
- What we have estimated is a *distribution* over the states
- If we had to guess ***a*** state, we would pick the most likely state from the distributions
- State(T=0) = Accelerating



- State(T=1) = Cruising

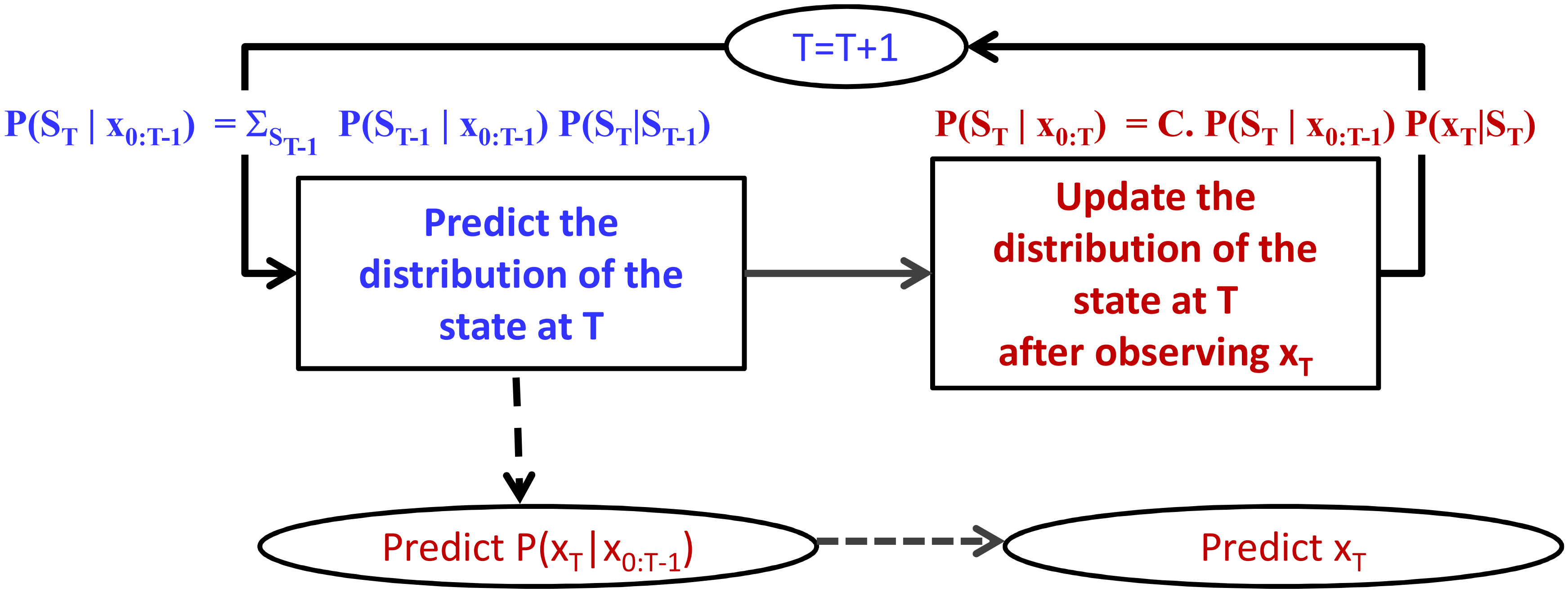


Estimating the *state*



- 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:
  - $\mathrm{argmax}_{x_T} P(x_T|x_{0:T-1})$
- MMSE estimate:
  - $\mathrm{Expectation}(x_T|x_{0:T-1})$

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}

A *continuous* state model

- HMM assumes a very coarsely quantized state space
  - Idling / accelerating / cruising / decelerating
- Actual state can be finer
  - Idling, accelerating at various rates, decelerating at various rates, cruising at various speeds
- Solution: Many more states (one for each acceleration /deceleration rate, crusing speed)?
- Solution: A *continuous* valued state

Tracking and Prediction: The wind and the target

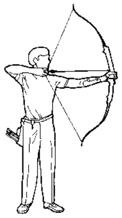
- Aim: measure wind velocity
- Using a noisy wind speed sensor
  - E.g. arrows shot at a target

- **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

$$Y_t = AS_t + \gamma_t$$



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

States are still "hidden"

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

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

- The state is a continuous valued parameter that is not directly seen
  - The state is the position of the automobile or the star
- The observations are dependent on the state and are the only way of knowing about the state
  - Sensor readings (for the automobile) or recorded image (for the telescope)

Statistical Prediction and Estimation

- Given an *a priori* probability distribution for the state
  - $P_0(s)$: Our belief in the state of the system before we observe any data
    - Probability of state of navlab
    - Probability of state of stars
- Given a sequence of observations $o_0..o_t$
- Estimate state at time t

Prediction and update at t = 0

- Prediction
  - Initial probability distribution for state
  - $P(s_0) = P_0(s_0)$
- Update:
  - Then we observe o_0
  - We must update our belief in the state

$$P(s_0 \mid o_0) = \frac{P(s_0)P(o_0 \mid s)}{P(o_0)} = \frac{P_0(s_0)P(o_0 \mid s_0)}{P(o_0)}$$

- $P(s_0|o_0) = C.P_0(s_0)P(o_0|s_0)$

Prediction and update at t = 0

- Prediction
  - Initial probability distribution for state
  - $P(s_0) = P_0(s_0)$
- Update:
  - Then we observe o_0
  - We must update our belief in the state

$$P(s_0 | o_0) = \frac{P(s_0)P(o_0 | s)}{P(o_0)} = \frac{P_0(s_0)P(o_0 | s_0)}{P(o_0)}$$

- $P(s_0|o_0) = C.P_0(s_0)P(o_0|s_0)$

The observation probability: P(o|s)

- $o_t = g(s_t, \gamma_t)$
  - This is a (possibly many-to-one) stochastic function of state s_t and noise γ_t
  - Noise γ_t is random. Assume it is the same dimensionality as o_t

- Let $P_\gamma(\gamma_t)$ be the probability distribution of γ_t
- Let $\{\gamma : g(s_t, \gamma) = o_t\}$ be all γ that result in o_t

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

The observation probability

- $P(o|s) = ?$

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

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

- The J is a Jacobian

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

- For scalar functions of scalar variables, it is simply a derivative: $|J_\gamma(g(s_t,\gamma))| = \left|\frac{\partial o_t}{\partial \gamma}\right|$

Predicting the next state at t=1

- Given $P(s_0|o_0)$, what is the probability of the state at t=1

$$P(s_1 | o_0) = \int_{\{s_0\}} P(s_1, s_0 | o_0) ds_0 = \int_{\{s_0\}} P(s_1 | s_0) P(s_0 | o_0) ds_0$$

- State progression function:

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

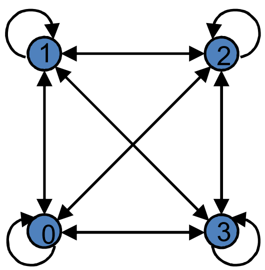
  - ε_t is a driving term with probability distribution $P_\varepsilon(\varepsilon_t)$

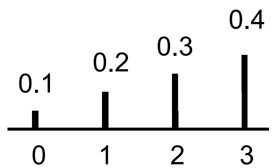
- $P(s_t|s_{t-1})$ can be computed similarly to $P(o|s)$
  - $P(s_1|s_0)$ is an instance of this

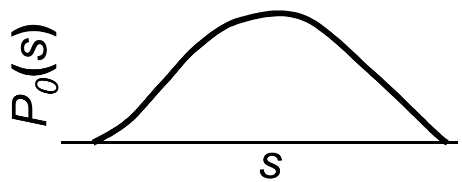
And moving on

- $P(s_1|o_0)$ is the predicted state distribution for t=1
- Then we observe o_1
  - We must update the probability distribution for s_1
  - $P(s_1|o_{0:1}) = CP(s_1|o_0)P(o_1|s_1)$
- We can continue on

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:

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

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

Update after O_0:

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

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

Prediction at time 1:

$$P(S_1|O_0) = \sum_{S_0} P(S_0|O_0)P(S_1|S_0)$$

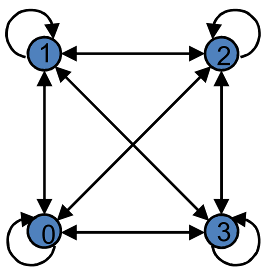
$$P(S_1|O_0) = \int_{-\infty}^{\infty} P(S_0|O_0)P(S_1|S_0)dS_0$$

Update after O_1:

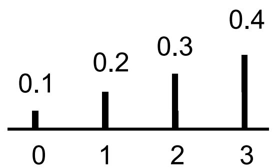
$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1)$$

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

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:

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

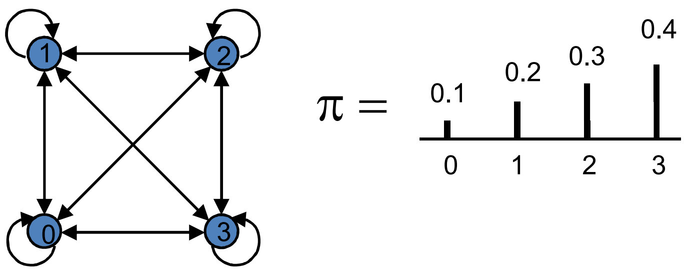
$$P(S_t|O_{0:t-1}) = \int_{-\infty}^{\infty} P(S_{t-1}|O_{0:t-1}) P(S_t|S_{t-1}) dS_{t-1}$$

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



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

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

Parameters

	Discrete	Continuous
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

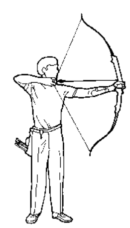
Linear model example
The wind and the target

- **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

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

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

- 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

$$o_t = B_t s_t + \gamma_t$$

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

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

- 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

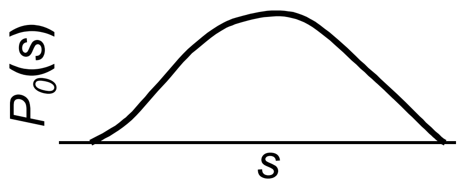
$$s_{t+1} = A_t s_t + \varepsilon_t$$

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

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

- 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:

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

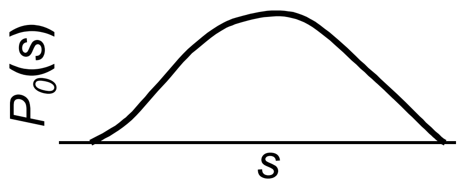
Prediction at time 1:

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

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:

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

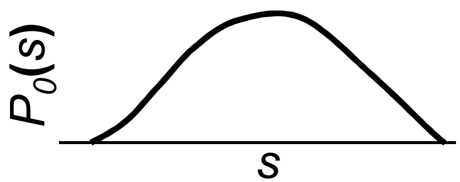
Model Parameters: The initial state probability

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

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

- 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:

$$P(S_0) = P_0(S_0) \qquad = N(\bar{s}_0, R_0)$$

Update after O_0:

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

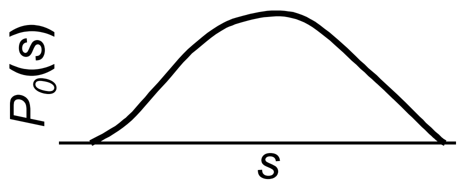
Prediction at time 1:

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

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:

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

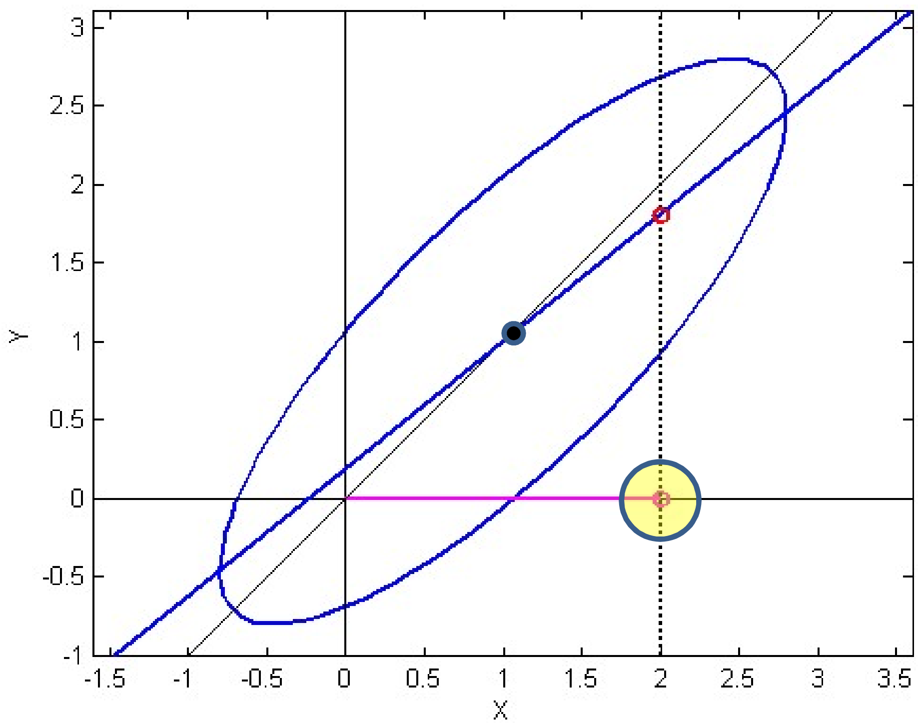
Prediction at time 1:

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

Update after O_1:

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

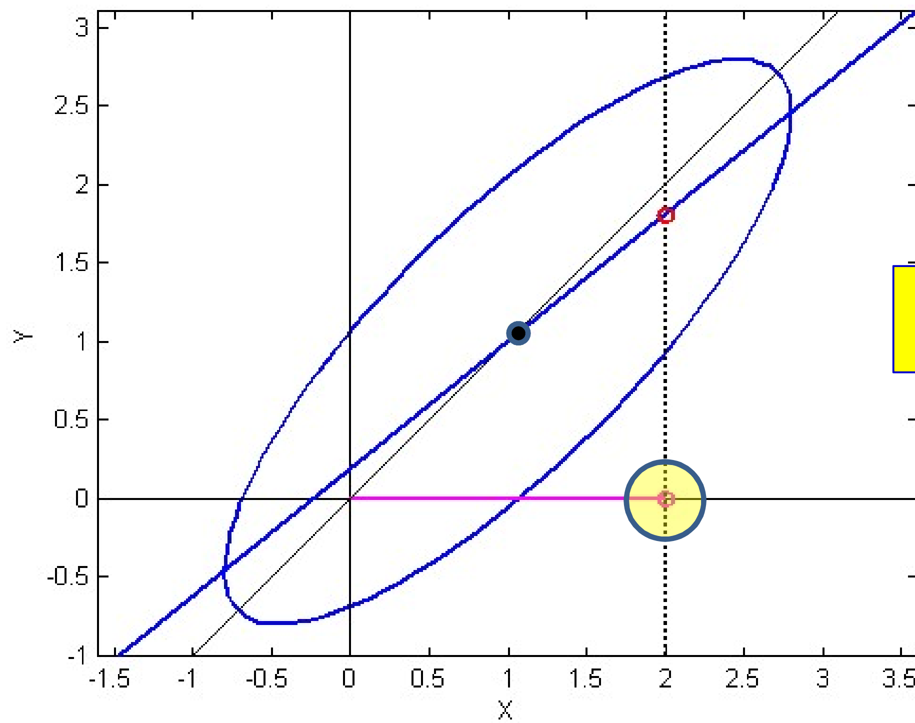
Recap: Conditional of S given O: P(S|O) for Gaussian RVs



$$O = BS + \gamma$$

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

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



$$O = BS + \gamma$$

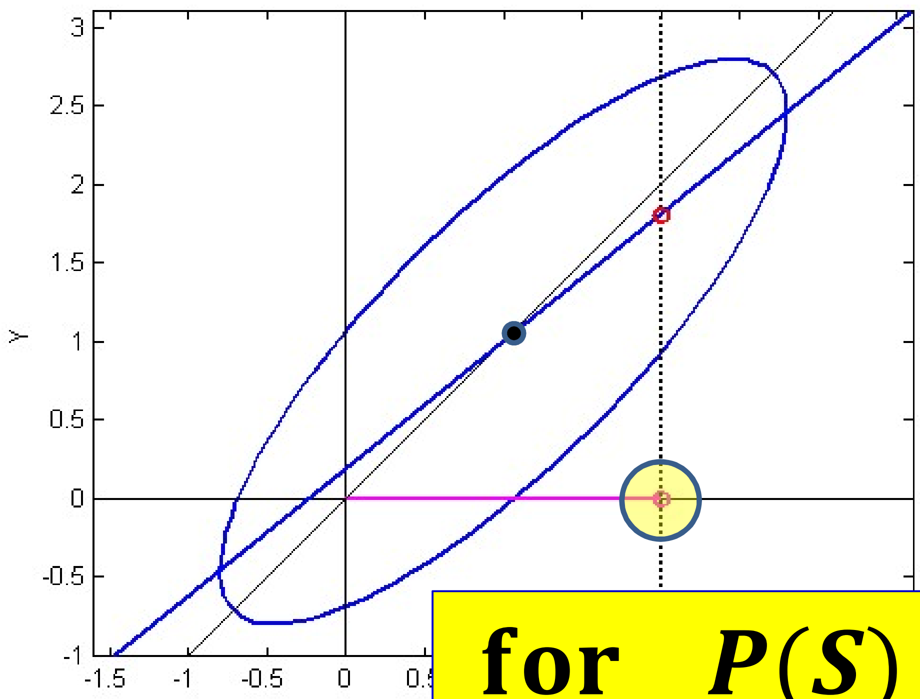
$$\Theta_{SO} = \Theta_S B^{\mathrm{T}}$$

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

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

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

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

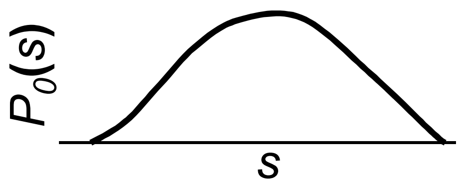


$$O = BS + \varepsilon$$

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

$$P(S_0|O_0) = N(\bar{s}_0 + R_0B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}(O_0 - B\bar{s}_0 - \mu_\gamma), \\ R_0 - R_0B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}BR_0)$$

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:

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

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

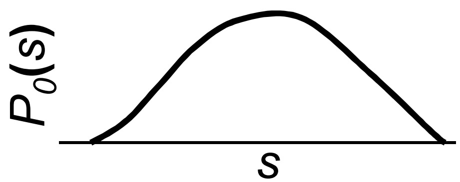
Prediction at time 1:

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

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:

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

Update after O_0:

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

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

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

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

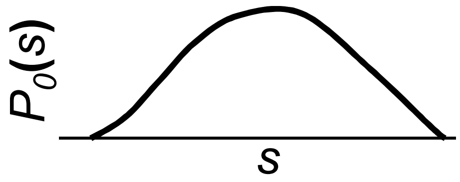
Prediction at time 1:

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

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:

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

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

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:

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

Introducting shorthand notation

$$P(S_0|O_0) = N(\bar{s}_0 + R_0B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}(O_0 - B\bar{s}_0 - \mu_\gamma),$$
$$R_0 - R_0B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}BR_0)$$

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

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

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

Introducting shorthand notation

$$P(S_0|O_0) = N(\bar{s}_0 + R_0B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}(O_0 - B\bar{s}_0 - \mu_\gamma),$$
$$R_0 - R_0B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}BR_0)$$

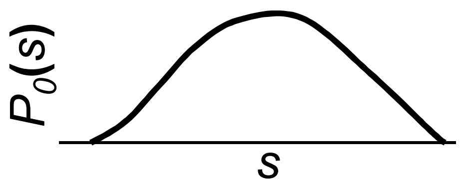
$$\boldsymbol{K_0 = R_0B^{\mathrm{T}}\left(BR_0B^{\mathrm{T}} + \Theta_\gamma\right)^{-1}}$$

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

$$\boldsymbol{\hat{R}_0 = (I - K_0B)R_0}$$

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

Continuous state systems



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

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

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

Update after O_0:

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

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

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

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

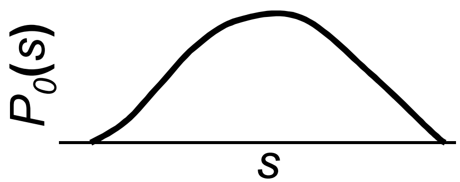
Prediction at time 1:

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

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

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

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

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

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

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

- 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!

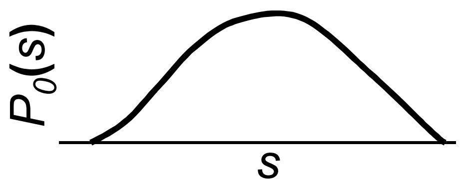
$$P(S_1|O_0) = \int_{-\infty}^{\infty} Gaussian\left(S_0; \hat{s}_0, \hat{R}_0\right) Gaussian(S_1; AS_0 + \mu_\varepsilon, \Theta_\varepsilon) dS_0$$

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

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

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

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:

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

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

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

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

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

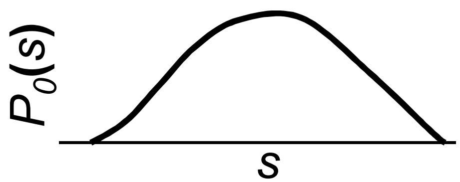
$$P(S_1|O_0) = N(A\hat{s}_0 + \mu_\varepsilon,\ \Theta_\varepsilon + A\hat{R}_0A^T)$$

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

$$R_1 = \Theta_\varepsilon + A\hat{R}_0A^T$$

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

Continuous state systems



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

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

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

Update after O_0:

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

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

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

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

Prediction at time 1:

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

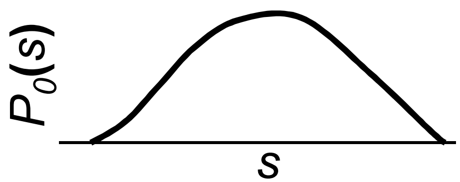
$$\bar{s}_1 = A\hat{s}_0 + \mu_\varepsilon$$

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

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:

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

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

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

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

Prediction at time 1:

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

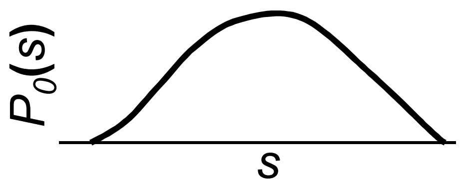
$$P(S_1|O_0) = N(\bar{s}_1, R_1)$$

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

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:

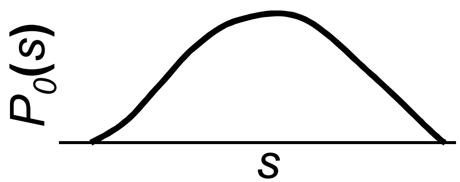
$$P(S_1|O_{0:1}) = C.P(S_1|O_0)P(O_1|S_1) = N(\hat{\boldsymbol{s}}_1, \hat{\boldsymbol{R}}_1)$$

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

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

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

Continuous state systems



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

$$o_t = B_t s_t + \gamma_t$$

Prediction at time 0:

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

Update after O_0:

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

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

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

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

Prediction at time 1:

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

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

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

Update after O_1:

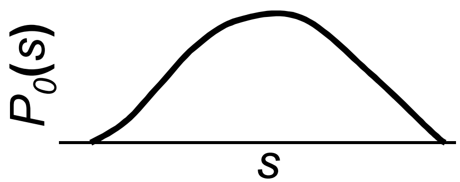
$$P(S_1|O_{0:1}) = N(\hat{s}_1, \hat{R}_1)$$

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

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

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

Gaussian Continuous State Linear Systems



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

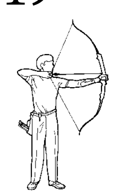
$$o_t = B_t s_t + \gamma_t$$

Prediction at time t:

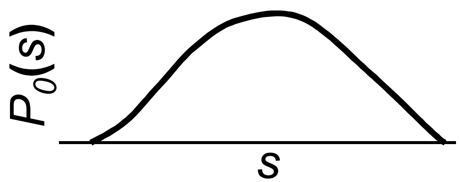
$$P(S_t|O_{0:t-1}) = \int_{-\infty}^{\infty} P(S_{t-1}|O_{0:t-1}) P(S_t|S_{t-1}) dS_{t-1}$$

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:

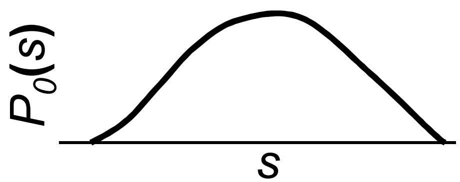
$$P(S_t|O_{0:t}) = N(\hat{s}_t, \hat{R}_t)$$

$$K_t = R_1 B^T \left(BR_1B^T + \Theta_\gamma\right)^{-1}$$

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

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

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

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

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

$$K_t = R_1 B^T \left(BR_1B^T + \Theta_\gamma\right)^{-1}$$

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

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

The Kalman filter

- Prediction (based on state equation)

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

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

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

- Update (using observation and observation equation)

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

$$o_t = B_t s_t + \gamma_t$$

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

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

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

$$s_t = s_t + K_t(o_t - B_t s_t - \mu_\gamma)$$

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

The Kalman filter

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

$$o_t = B_t s_t + \gamma_t$$

- Prediction

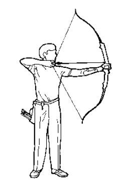
$$\bar{s}_t = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

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

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

- Update

Actual observation



$$o_t$$

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

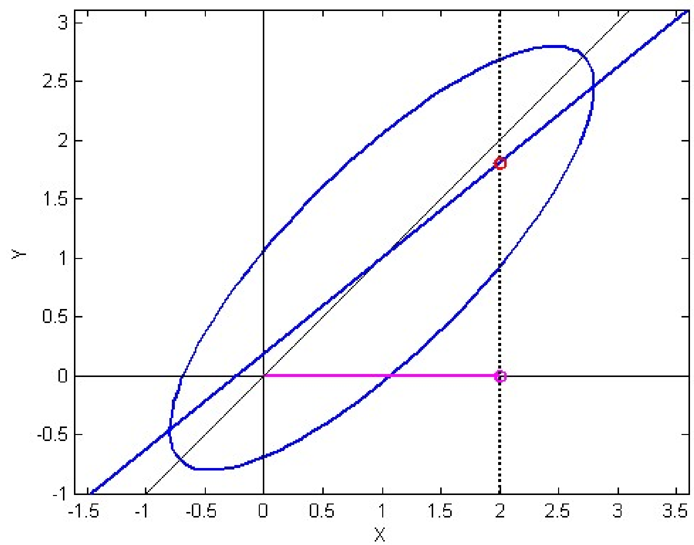
$$\hat{s}_t = \bar{s}_t + K_t \left(o_t - B_t \bar{s}_t \right)$$

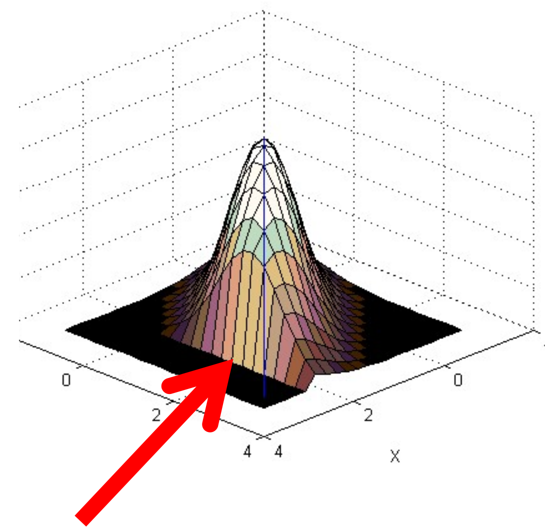
$$\hat{R}_t = \left(I - K_t B_t \right) R_t$$

MAP Recap (for Gaussians)

- If P(x,y) is Gaussian:

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





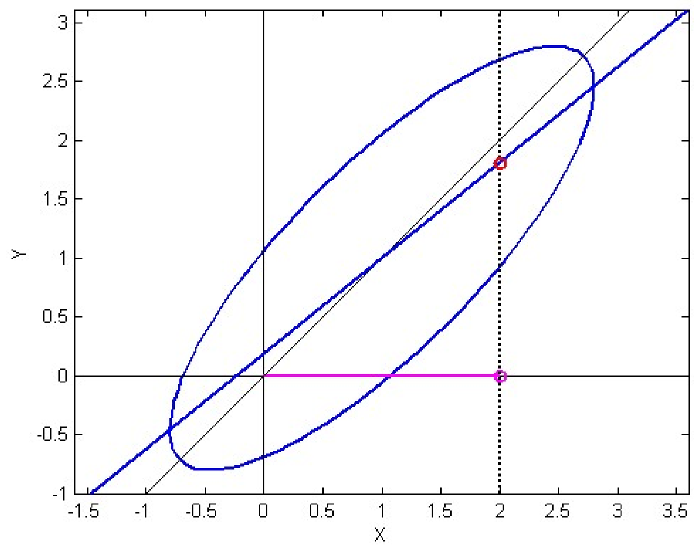
$$P(y\,|\,x) = N(\mu_y + C_{yx}C_{xx}^{-1}(x-\mu_x), C_{yy} - C_{yx}^T C_{xx}^{-1} C_{xy})$$

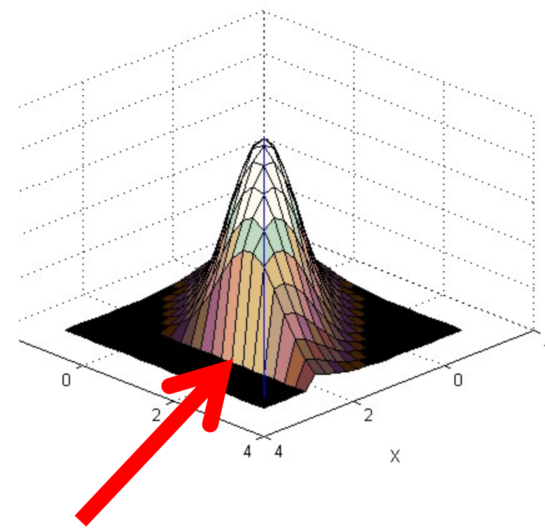
$$\hat{y} = \mu_y + C_{yx}C_{xx}^{-1}(x-\mu_x)$$

MAP Recap: For Gaussians

- If P(x,y) is Gaussian:

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





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

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

“Slope” of the line

The Kalman filter

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

$$o_t = B_t s_t + \gamma_t$$

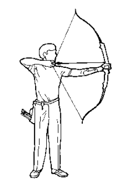
- Prediction

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

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

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

- Update



$$o_t$$

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

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**

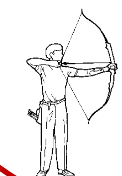
The Kalman filter

- Prediction

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

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

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



$$o_t$$

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

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

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

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

The Kalman filter

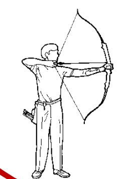
$$s_t = A_t s_{t-1} + \varepsilon_t$$

- Prediction

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

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

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

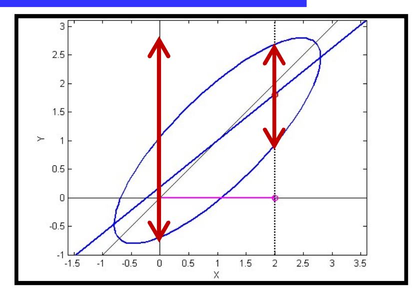
$$R_t = \Theta_\varepsilon + A_t \hat{R}_{t-1} A_t^T$$

- 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

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



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:

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

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

- Update

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

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

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

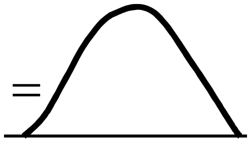
$$o_t = g(s_t, \gamma_t)$$

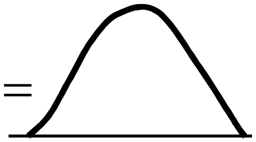
- f() and/or g() may not be nice linear functions
  - Conventional Kalman update rules are no longer valid
- ε and/or γ may not be Gaussian
  - Gaussian based update rules no longer valid

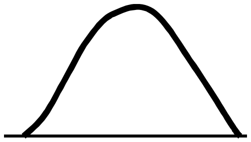
Linear Gaussian Model

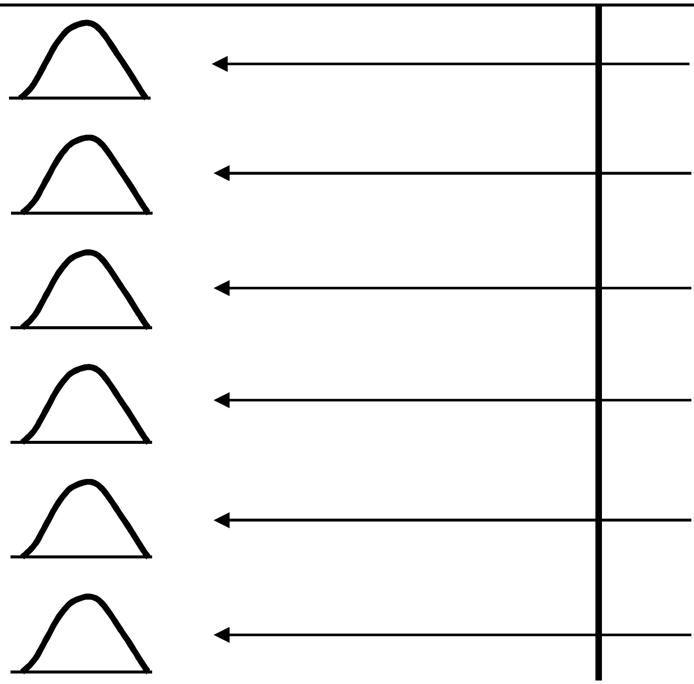
$$s_t = A_t s_{t-1} + \varepsilon_t$$

$$o_t = B_t s_t + \gamma_t$$

$P(s)$ =  *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

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

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

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

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

$$P(s_t \mid \mathrm{o}_{0:t}) = CP(s_t \mid \mathrm{o}_{0:t-1}) P(\mathrm{o}_t \mid s_t)$$

- 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

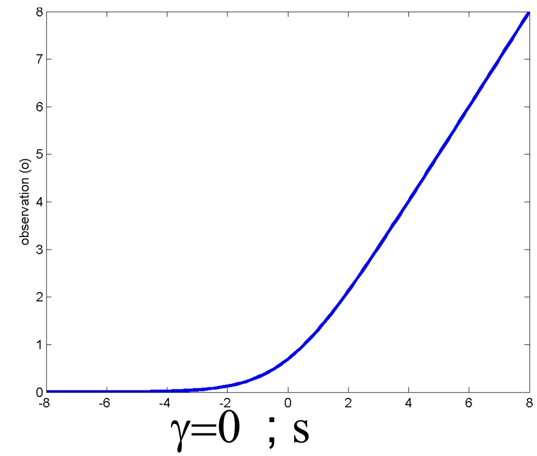
$$P(o_t \mid s_t) = \sum_{\gamma : g(s_t,\gamma) = o_t} \frac{P_\gamma(\gamma)}{|J_{g(s_t,\gamma)}(o_t)|}$$

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

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

Example: a simple nonlinearity

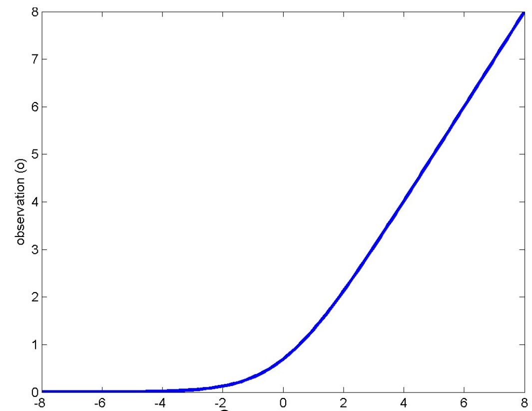
$$o = \gamma + \log(1 + \exp(s))$$



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

Example: a simple nonlinearity

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



γ=0 ; s

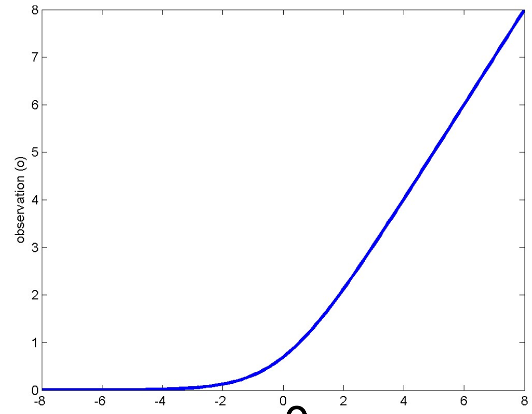
- P(o|s) = ?

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

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

Example: At T=0.

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



- Assume initial probability P(s) is Gaussian

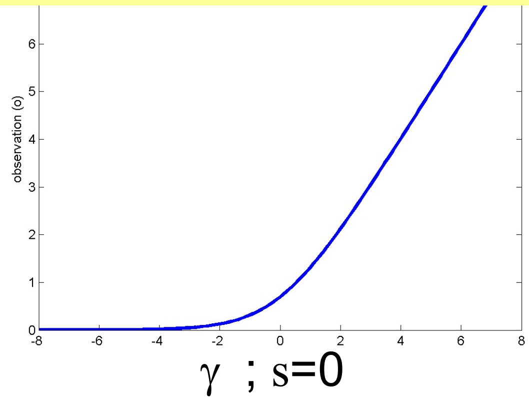
$$P(s_0) = P_0(s) = Gaussian(s; \bar{s}, R)$$

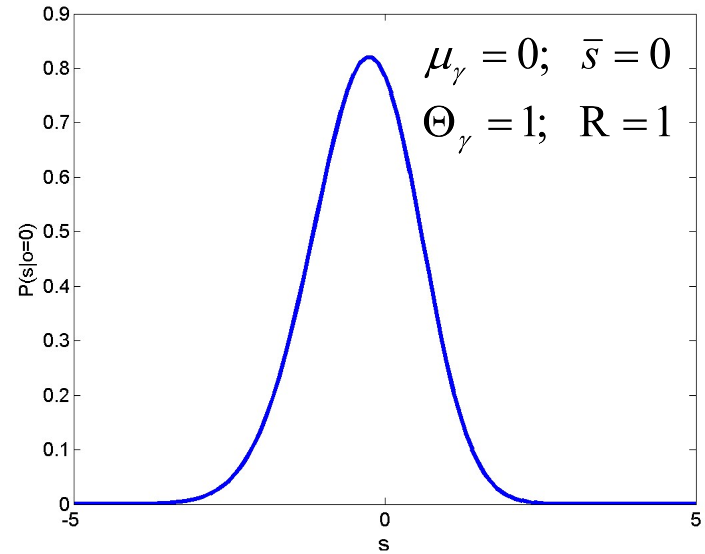
- Update $$P(s_0 \mid o_0) = CP(o_0 \mid s_0)P(s_0)$$

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

UPDATE: At T=0.

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





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

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

- = 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 $P(s_1 \mid \mathrm{o}_0) = \int_{-\infty}^{\infty} P(s_0 \mid \mathrm{o}_0) P(s_1 \mid s_0) ds_0$

$$P(s_1 \mid \mathrm{o}_0) = \int_{-\infty}^{\infty} C \exp\begin{pmatrix} -0.5(\mu_\gamma + \log(1+\exp(s_0)) - o)^T \Theta_\gamma^{-1} (\mu_\gamma + \log(1+\exp(s_0)) - o) \\ -0.5(s_0 - \bar{s})^T R^{-1} (s_0 - \bar{s}) \end{pmatrix} \exp\left((s_1 - s_0)^T \Theta_\varepsilon^{-1} (s_1 - s_0)\right) ds_0$$

- = 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

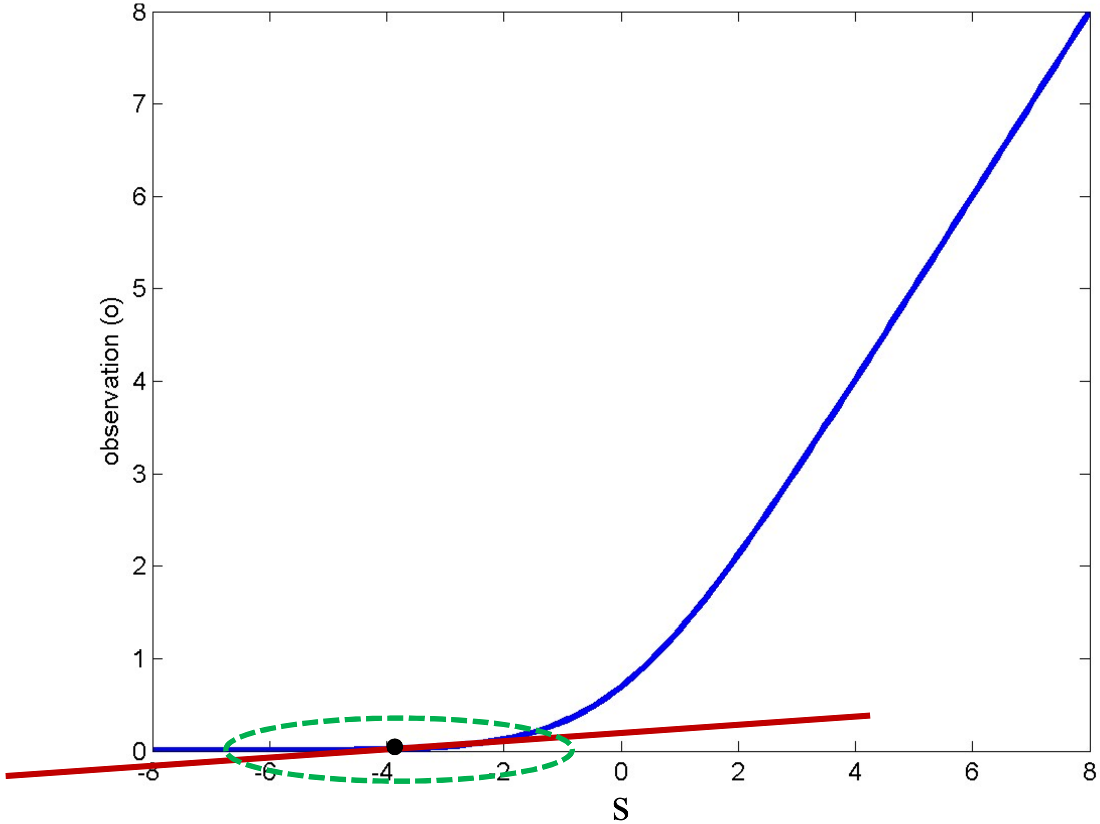
The State prediction Equation

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

- 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

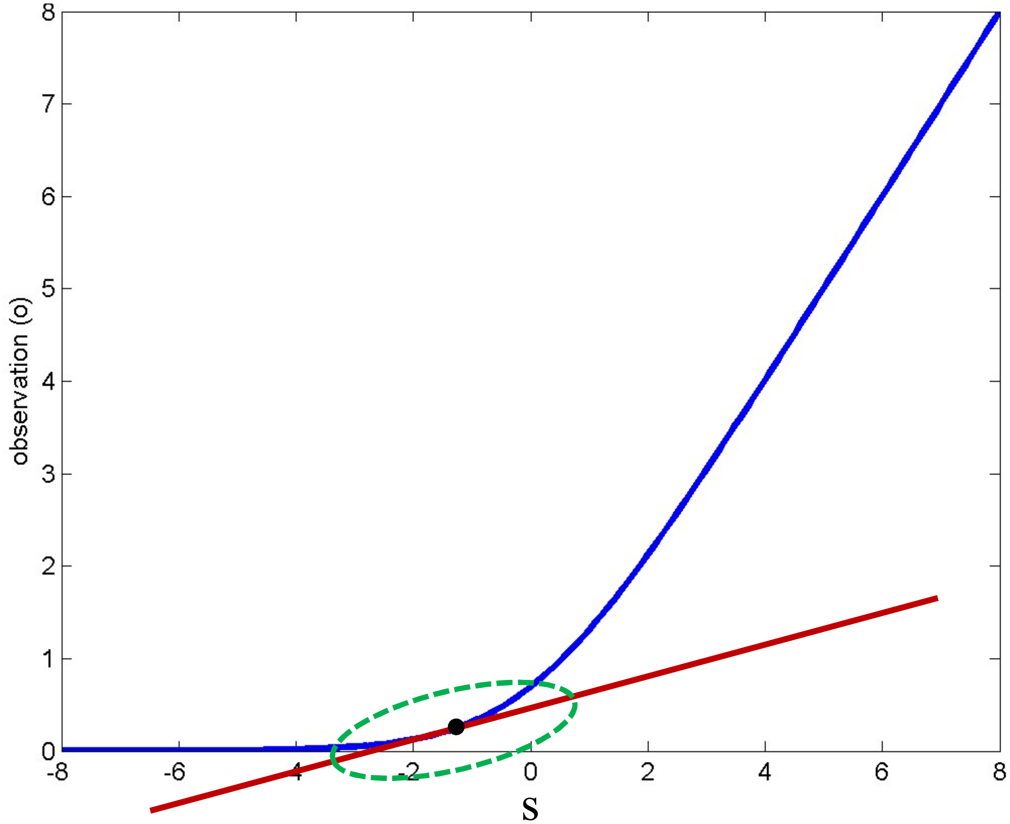
Simplifying the problem: Linearize



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

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

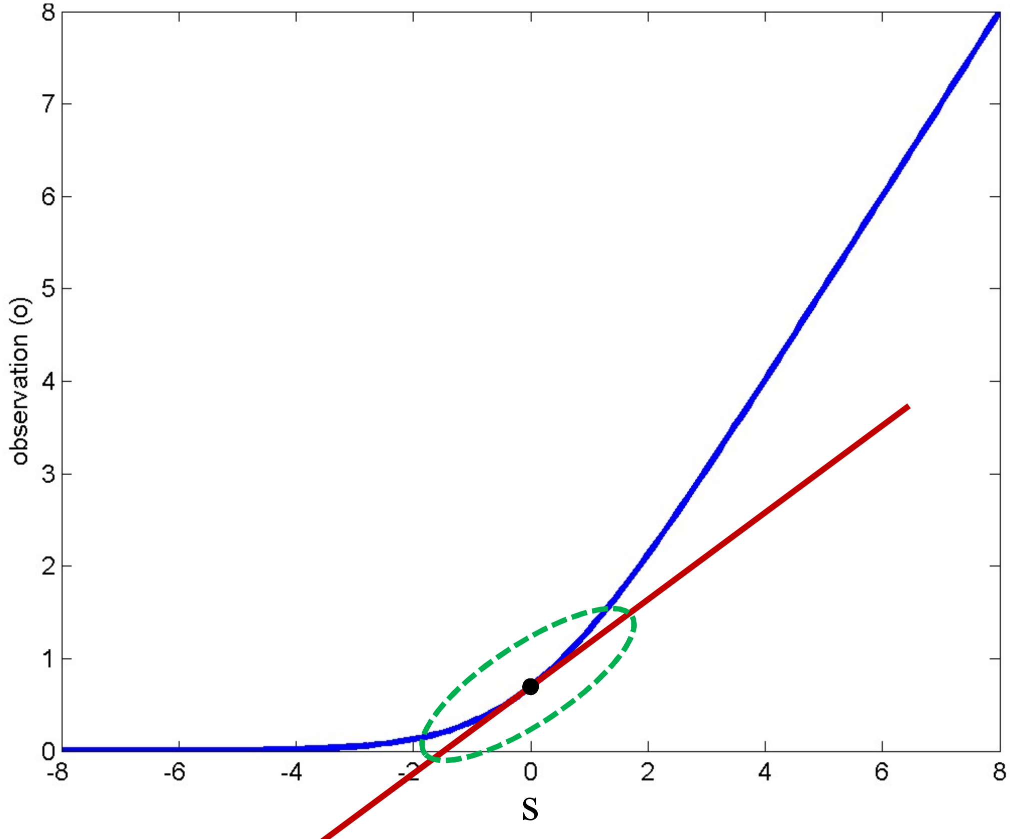
Simplifying the problem: Linearize



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

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

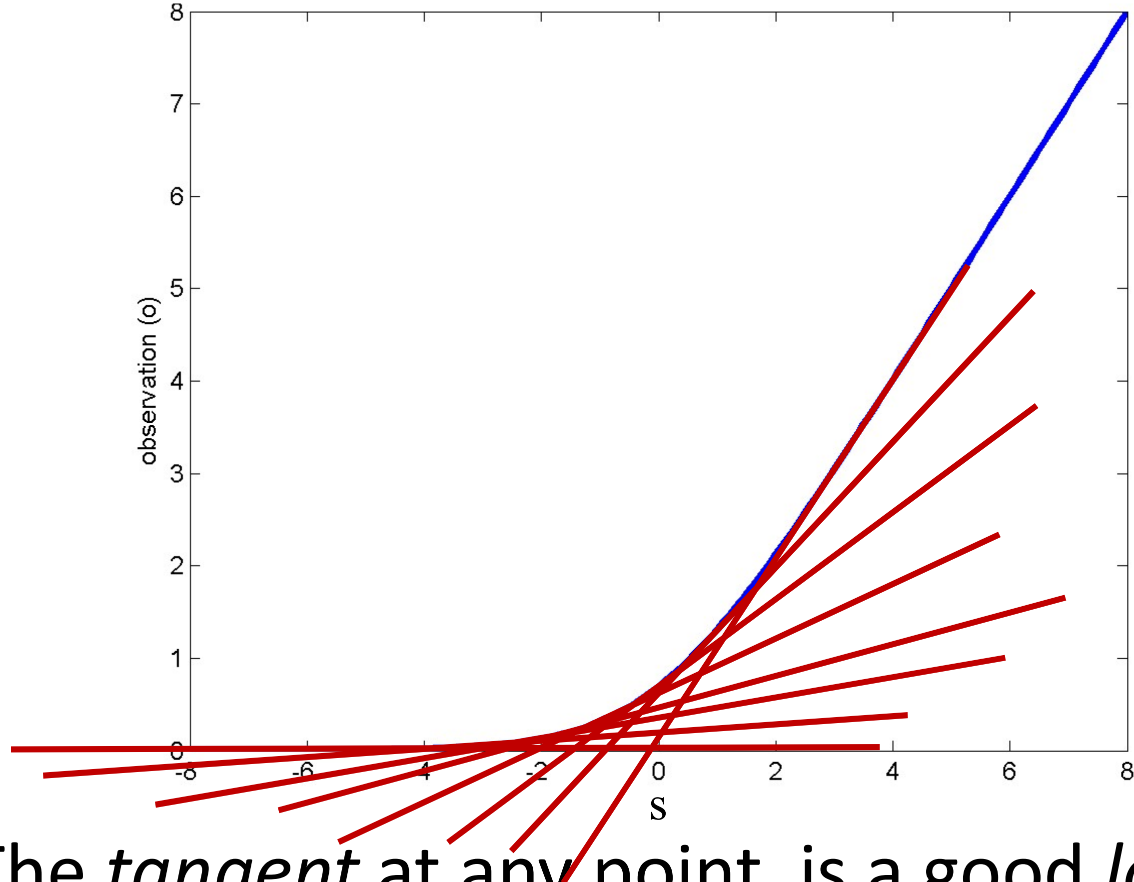
Simplifying the problem: Linearize



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

- 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

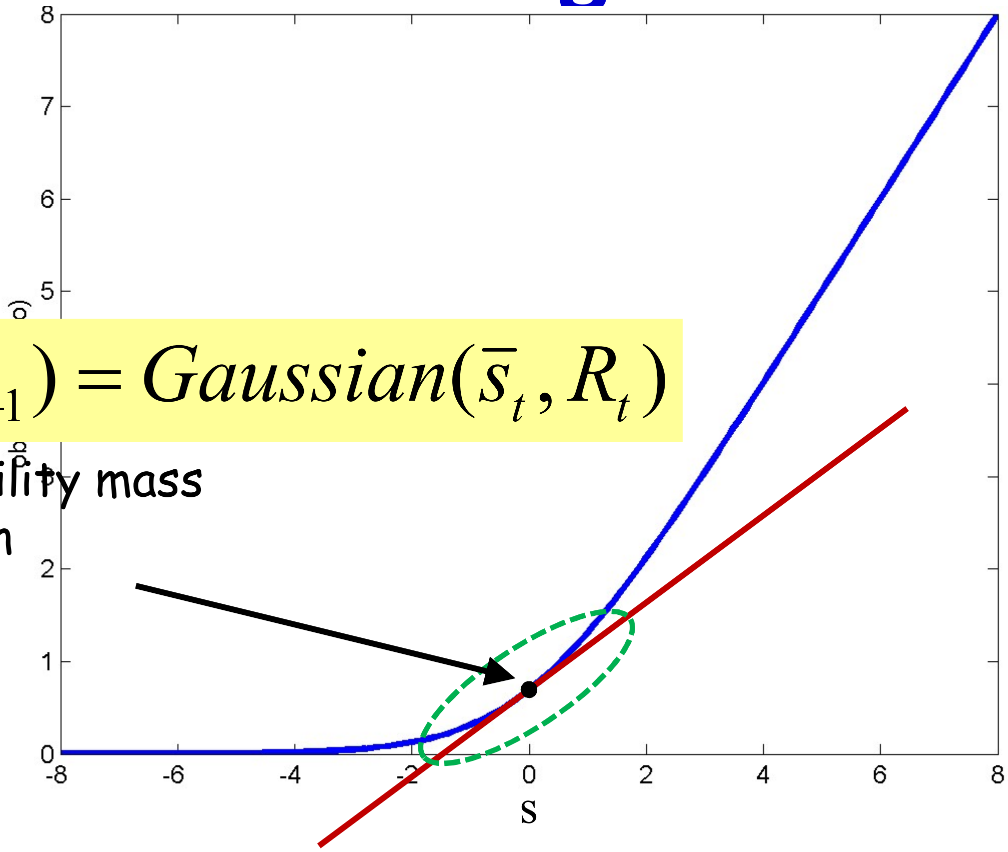
Linearizing the observation function

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

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

- 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

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

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

- 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

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

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

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

Assuming ε and γ are 0 mean for simplicity

The Extended Kalman filter

- Prediction

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

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

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

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

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

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

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

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

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()

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

The Extended Kalman filter

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

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

- Prediction

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

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

- Update

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

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

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

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

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

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

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

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

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

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

- Prediction

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

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

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

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

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

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

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

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

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

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

- Prediction

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

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

- Update

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

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

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

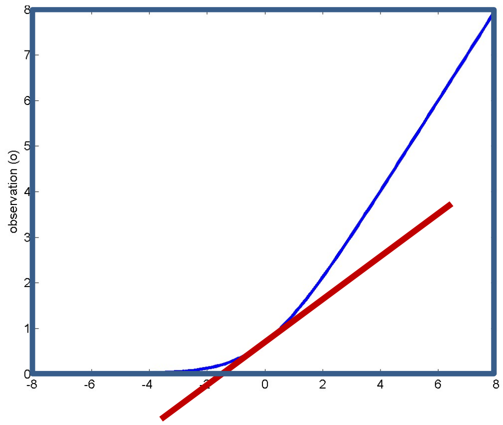
EKFs

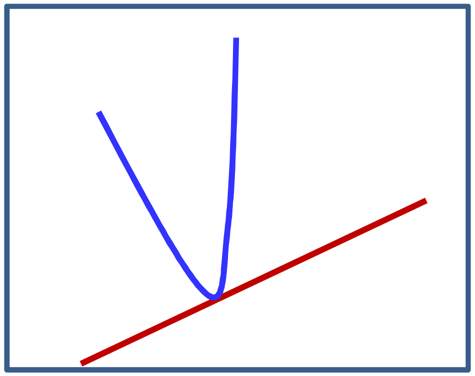
- EKFs are probably the most commonly used algorithm for tracking and prediction
  – Most systems are non-linear
  – Specifically, the relationship between state and observation is usually nonlinear
  – The approach can be extended to include non-linear functions of noise as well

- 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