11-755 Machine Learning for Signal Processing

Predicting and Estimation from Time Series

Class 16. 25 Oct 2012

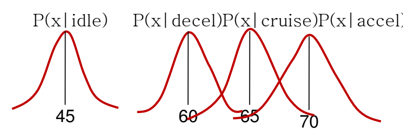
An automotive example

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

What we know

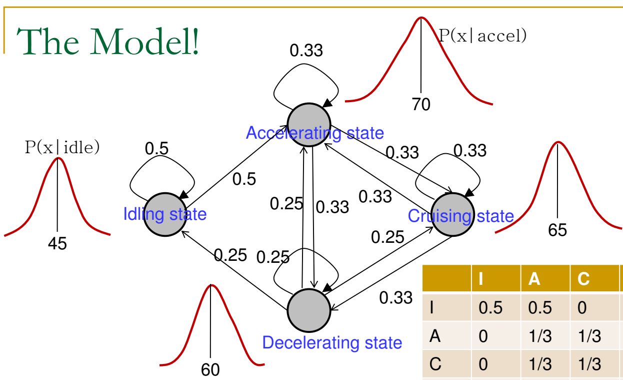
- An automobile that is at rest can accelerate, or continue to stay at rest
- An accelerating automobile can hit a steady-state velocity, continue to accelerate, or decelerate
- A decelerating automobile can continue to decelerate, come to rest, cruise, or accelerate
- A automobile at a steady-state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

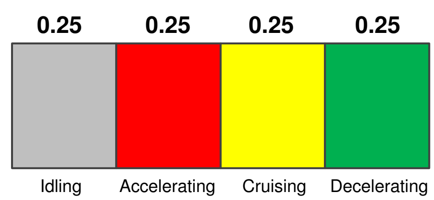
The Model!



	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

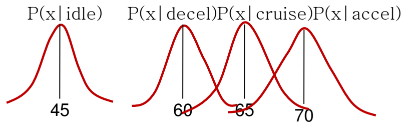
- The state-space model
  - Assuming all transitions from a state are equally probable

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

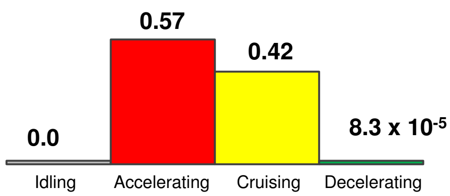


- At T=0 we observe the sound level x_0 = 67dB SPL
  - The observation modifies our belief in the state of the system
- $P(x_0|idle) = 0$
- $P(x_0|deceleration) = 0.0001$
- $P(x_0|acceleration) = 0.7$
- $P(x_0|cruising) = 0.5$
  - Note, these don't have to sum to 1
  - In fact, since these are densities, any of them can be > 1

Estimating state after at observing x_0

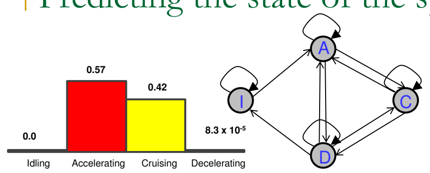
- $P(state \mid x_0) = C\, P(state)P(x_0|state)$
  - $P(idle \mid x_0) = 0$
  - $P(deceleration \mid x_0) = C\ 0.000025$
  - $P(cruising \mid x_0) = C\ 0.125$
  - $P(acceleration \mid x_0) = C\ 0.175$
- Normalizing
  - $P(idle \mid x_0) = 0$
  - $P(deceleration \mid x_0) = 0.000083$
  - $P(cruising \mid x_0) = 0.42$
  - $P(acceleration \mid x_0) = 0.57$

Estimating the state at T = 0+



- At T=0, after the first observation, we must 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

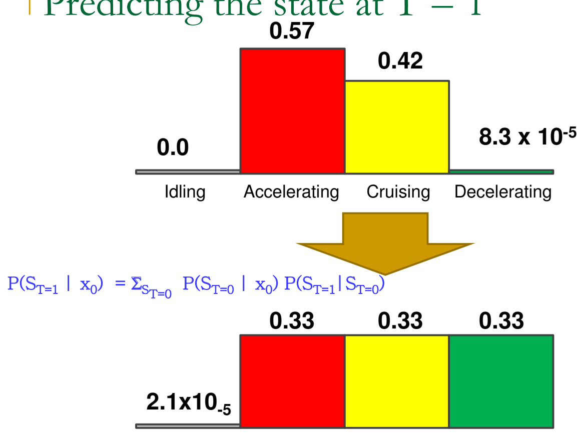
Predicting the state of the system at T=1



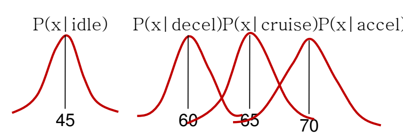
	I	A	C	D
I	0.5	0.5	0	0
A	0	1/3	1/3	1/3
C	0	1/3	1/3	1/3
D	0.25	0.25	0.25	0.25

- 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
  - $P(S_{T=1} \mid x_0) = \Sigma_{S_{T=0}} P(S_{T=0} \mid x_0)\, P(S_{T=1}|S_{T=0})$

Predicting the state at T = 1



Updating after the observation at T=1

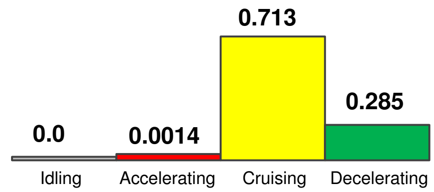


- At T=1 we observe x_1 = 63dB SPL
- $P(x_1|idle) = 0$
- $P(x_1|deceleration) = 0.2$
- $P(x_1|acceleration) = 0.001$
- $P(x_1|cruising) = 0.5$

Update after observing x_1

- $P(\text{state} \mid x_{0:1}) = C\, P(\text{state} \mid x_0) P(x_1 \mid \text{state})$
  - $P(\text{idle} \mid x_{0:1}) = 0$
  - $P(\text{deceleration} \mid x_{0,1}) = C\ 0.066$
  - $P(\text{cruising} \mid x_{0:1}) = C\ 0.165$
  - $P(\text{acceleration} \mid x_{0:1}) = C\ 0.00033$
- Normalizing
  - $P(\text{idle} \mid x_{0:1}) = 0$
  - $P(\text{deceleration} \mid x_{0:1}) = 0.285$
  - $P(\text{cruising} \mid x_{0:1}) = 0.713$
  - $P(\text{acceleration} \mid x_{0:1}) = 0.\ 0014$

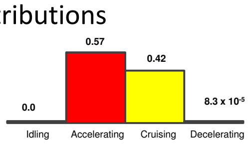
Estimating the state at T = 1+



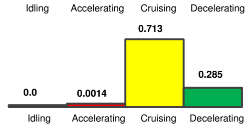
- The updated probability at T=1 incorporates information from both x_0 and x_1
  - It is NOT a local decision based on x_1 alone
  - Because of the Markov nature of the process, the state at T=0 affects the state at T=1
    - x_0 provides evidence for the state at T=1

Estimating a Unique state

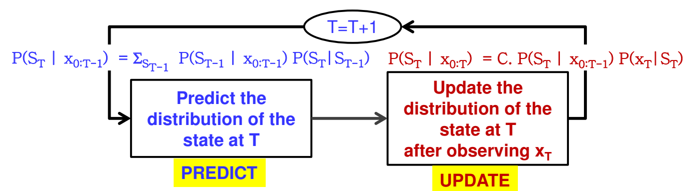
- What we have estimated is a *distribution* over the states
- If we had to guess ***a*** state, we would pick the most likely state from the distributions
- State(T=0) = Accelerating



- State(T=1) = Cruising

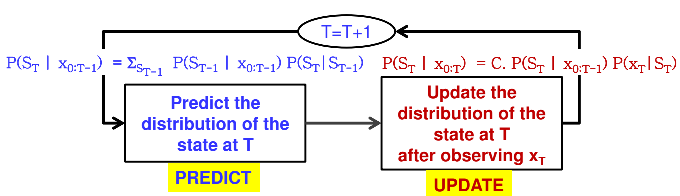


Overall procedure



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

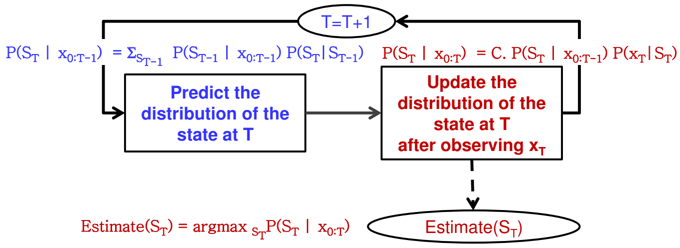
    PREDICT

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

Decomposing the forward algorithm

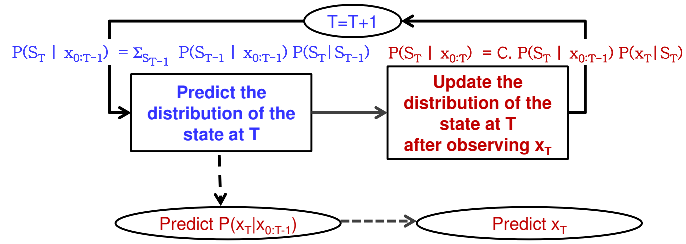
- $P(x_{0:T},S_T) = P(x_T \mid S_T)\ \Sigma_{S_{T-1}}\ P(x_{0:T-1}, S_{T-1})\ P(S_T \mid S_{T-1})$
- Predict:
- $P(x_{0:T-1},S_T) = \Sigma_{S_{T-1}}\ P(x_{0:T-1}, S_{T-1})\ P(S_T \mid S_{T-1})$
- Update:
- $P(x_{0:T},S_T) = P(x_T \mid S_T)\ P(x_{0:T-1},S_T)$

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:
  - $P(x_T|x_{0:T-1}) = \Sigma_{S_T} P(x_T|S_T)\, P(S_T|x_{0:T-1})$
- The actual observation can be predicted from $P(x_T|x_{0:T-1})$

Predicting the next observation

- 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}

A *known* 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

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

Continuous state system

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

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

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

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

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

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 the set of γ that result in o_t

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

The observation probability

- P(o|s) = ? $o_t = g(s_t, \gamma_t)$

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

- The J is a jacobian

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

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

Predicting the next state

- 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:

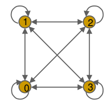
$$s_t = f(s_{t-1}, \varepsilon_t)$$

  - 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

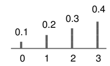
And moving on

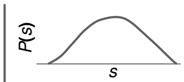
- $P(s_1|o_0)$ is the predicted state distribution for t=1
- Then we observe o_1
  - We must update the probability distribution for s1
  - $P(s_1|o_{0:1}) = CP(s_1|o_0)P(o_1|s_1)$
- We can continue on

Discrete vs. Continuous state systems



$\pi =$



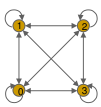


$s_t = f(s_{t-1}, \varepsilon_t)$

$o_t = g(s_t, \gamma_t)$

Prediction at time 0: $P(s_0) = \pi(s_0)$	$P(s_0) = P(s)$
Update after O_0: $P(s_0 \mid O_0) = C\,\pi(s_0)P(O_0 \mid s_0)$	$P(s_0 \mid O_0) = C\,P(s_0)\,P(O_0 \mid s_0)$
Prediction at time 1: $P(s_1 \mid O_0) = \sum_{s_0} P(s_0 \mid O_0)P(s_1 \mid s_0)$	$P(s_1 \mid O_0) = \int_{-\infty}^{\infty} P(s_0 \mid O_0)P(s_1 \mid s_0)ds_0$
Update after O_1: $P(s_1 \mid O_0,O_1) = C\,P(s_1 \mid O_0)\,P(O_1 \mid s_1)$	$P(s_1 \mid O_0,O_1) = C\,P(s_1 \mid O_0)\,P(O_1 \mid s_1)$

Discrete vs. Continuous State Systems



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

$o_t = g(s_t, \gamma_t)$

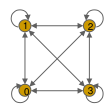
Prediction at time t

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

Update after O_t:

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

Discrete vs. Continuous State Systems



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

$o_t = g(s_t, \gamma_t)$

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

Special case: Linear Gaussian model

$s_t = A_t s_{t-1} + \varepsilon_t$ $\qquad P(\varepsilon) = \frac{1}{\sqrt{(2\pi)^d \mid \Theta_\varepsilon \mid}} \exp\left(-0.5(\varepsilon - \mu_\varepsilon)^T \Theta_\varepsilon^{-1}(\varepsilon - \mu_\varepsilon)\right)$

$o_t = B_t s_t + \gamma_t$ $\qquad P(\gamma) = \frac{1}{\sqrt{(2\pi)^d \mid \Theta_\gamma \mid}} \exp\left(-0.5(\gamma - \mu_\gamma)^T \Theta_\gamma^{-1}(\gamma - \mu_\gamma)\right)$

- 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

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

The observation probability

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

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

- 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

The updated state probability at T=0

- $P(s_0 \mid o_0) = C\, P(s_0)\, P(o_0 \mid s_0)$

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

$$P(o_0 \mid s_0) = Gaussian(o_0; \mu_\gamma + B_0 s_0, \Theta_\gamma)$$

$$P(s_0 \mid o_0) = C Gaussian(s_0; \bar{s}, R) Gaussian(o_0; \mu_\gamma + B_0 s_0, \Theta_\gamma)$$

Note 1: product of two Gaussians

- The product of two Gaussians is a Gaussian

$$Gaussian(s; \bar{s}, R) Gaussian(o; \mu + Bs, \Theta)$$

$$C_1 \exp\left(-0.5(s-\bar{s})^T R^{-1}(s-\bar{s})\right) C_2 \exp\left(-0.5(o-\mu-Bs)^T \Theta^{-1}(o-\mu-Bs)\right)$$

$$C.Gaussian\left(s; \left(R^{-1} + B^T\Theta^{-1}B\right)^{-1}\left(R^{-1}\bar{s} + B^T\Theta^{-1}(o-\mu)\right), \left(R^{-1} + B^T\Theta^{-1}B\right)^{-1}\right)$$

Not a good estimate --

The updated state probability at T=0

- $P(s_0 \mid o_0) = C\, P(s_0)\, P(o_0 \mid s_0)$

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

$$P(o_0 \mid s_0) = Gaussian(o_0; \mu_\gamma + B_0 s_0, \Theta_\gamma)$$

$$P(s_0 \mid o_0) = Gaussian\left(s_0; \left(R^{-1} + B_0^T\Theta_\gamma^{-1}B_0\right)^{-1}\left(R^{-1}\bar{s} + B_0^T\Theta_\gamma^{-1}(o_0-\mu_\gamma)\right), \left(R^{-1} + B_0^T\Theta_\gamma^{-1}B_0\right)^{-1}\right)$$

$$P(s_0 \mid o_0) = Gaussian\left(s_0; \hat{s}_0, \hat{R}_0\right)$$

The state transition probability

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

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

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

Note 2: integral of product of two Gaussians

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

$$\int_{-\infty}^{\infty} Gaussian(x; \mu_x, \Theta_x) Gaussian(y; Ax + b, \Theta_y) dx =$$

$$\int_{-\infty}^{\infty} C_1 \exp\left(-0.5(x-\mu_x)^T \Theta_x^{-1}(x-\mu_x)\right) C_2 \exp\left(-0.5(y-Ax-b)^T \Theta_y^{-1}(y-Ax-b)\right) dx$$

$$= Gaussian\left(y; A\mu_x + b, \Theta_y + A\Theta_x A^T\right)$$

The predicted state probability at t=1

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

$$P(s_1 \mid s_0) = Gaussian(s_1; \mu_\varepsilon + A_1 s_0, \Theta_\varepsilon)$$

$$P(s_0 \mid o_0) = Gaussian\left(s_0; \hat{s}_0, \hat{R}_0\right)$$

$$P(s_1 \mid o_0) = \int_{-\infty}^{\infty} Gaussian\left(s_0; \hat{s}_0, \hat{R}_0\right) Gaussian(s_1; \mu_\varepsilon + A_1 s_0, \Theta_\varepsilon) ds_0$$

$$P(s_1 \mid o_0) = Gaussian\left(s_1; A_1\hat{s}_0 + \mu_\varepsilon, \Theta_\varepsilon + A_1\hat{R}_0 A_1^T\right)$$

- Remains Gaussian

The updated state probability at T=1

- $P(s_1 \mid o_{0:1}) = C\, P(s_1 \mid o_0)\, P(o_1 \mid s_1)$

$$P(s_1 \mid o_0) = Gaussian\left(s_1; A_1\hat{s}_0 + \mu_\varepsilon, \Theta_\varepsilon + A_1\hat{R}_0 A_1^T\right)$$

$$P(o_1 \mid s_1) = Gaussian(o_1; \mu_\gamma + B_1 s_1, \Theta_\gamma)$$

$$\vdots$$

$$P(s_1 \mid o_{0:1}) = Gaussian\left(s_1; \hat{s}_1, \hat{R}_1\right)$$

The Kalman Filter!

- Prediction at T

$$P(s_t \mid o_{0:t-1}) = Gaussian\left(s_t; A_t\hat{s}_{t-1} + \mu_\varepsilon, \Theta_\varepsilon + A_t\hat{R}_{t-1} A_t^T\right)$$

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

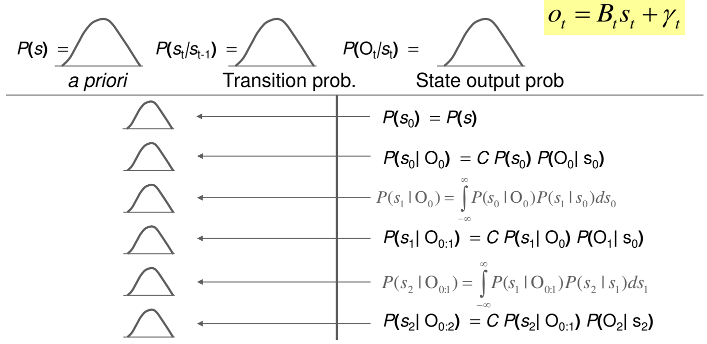
- Update at T

$$P(s_t \mid o_{0:t}) =$$
$$Gaussian\left(s_t; \left(R_t^{-1} + B_t^T\Theta_\gamma^{-1}B_t\right)^{-1}\left(R_t^{-1}\bar{s}_t + B_t^T\Theta_\gamma^{-1}(o_t - \mu_\gamma)\right), \left(R_t^{-1} + B_t^T\Theta_\gamma^{-1}B_t\right)^{-1}\right)$$

$$P(s_t \mid o_{0:t}) = Gaussian\left(s_t; \hat{s}_t, \hat{R}_t\right)$$

Linear Gaussian Model

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



All distributions remain Gaussian

The Kalman filter

- The actual state estimate is the *mean* of the updated distribution
- Predicted state at time *t*

$$\bar{s}_t = mean[P(s_t \mid o_{0:t-1})] = A_t\hat{s}_{t-1} + \mu_\varepsilon$$

- Updated estimate of state at time *t*

$$\hat{s}_t = mean[P(s_t \mid o_{0:t})] = \left(R_t^{-1} + B_t^T\Theta_\gamma^{-1}B_t\right)^{-1}\left(R_t^{-1}\bar{s}_t + B_t^T\Theta_\gamma^{-1}(o_t - \mu_\gamma)\right)$$

Stable Estimation

$$\hat{s}_t = mean[P(s_t \mid o_{0:t})] = \left(R_t^{-1} + B_t^T\Theta_\gamma^{-1}B_t\right)^{-1}\left(R_t^{-1}\bar{s}_t + B_t^T\Theta_\gamma^{-1}(o_t - \mu_\gamma)\right)$$

- The above equation fails if there is no observation noise
  - $\Theta_\gamma = 0$
  - Paradoxical?
  - Happens because we do not use the relationship between *o* and *s* effectively
- Alternate derivation required
  - Conventional Kalman filter formulation

A matrix inverse identity

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B\left(C - B^T A^{-1} B\right)^{-1} B^T A^{-1} & -A^{-1}B\left(C - B^T A^{-1}B\right)^{-1} \\ -\left(C - B^T A^{-1} B\right)^{-1} B^T A^{-1} & \left(C - B^T A^{-1} B\right)^{-1} \end{bmatrix}$$

- Work it out..

For any jointly Gaussian RV

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} \quad \mu_Z = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \quad C_Z = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{XY}^T & C_{YY} \end{bmatrix}$$

$$C_Z^{-1} = \begin{bmatrix} C_{XX}^{-1} + C_{XX}^{-1} C_{XY}\left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} C_{XY}^T C_{XX}^{-1} & -C_{XX}^{-1} C_{XY}\left(C - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} \\ -\left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} C_{XY}^T C_{XX}^{-1} & \left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} \end{bmatrix}$$

- Using the Matrix Inversion Identity

For any jointly Gaussian RV

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} \quad \mu_Z = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \quad C_Z = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{XY}^T & C_{YY} \end{bmatrix}$$

$$C_Z^{-1} = \begin{bmatrix} C_{XX}^{-1} + C_{XX}^{-1} C_{XY}\left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} C_{XY}^T C_{XX}^{-1} & -C_{XX}^{-1} C_{XY}\left(C - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} \\ -\left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} C_{XY}^T C_{XX}^{-1} & \left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} \end{bmatrix}$$

$$(Z - \mu_Z)^T C_Z^{-1} (Z - \mu_Z) = \quad Quadratic(X) +$$

$$\left(Y - \mu_Y - C_{YX} C_{XX}^{-1}(X - \mu_X)\right)^T \left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} \left(Y - \mu_Y - C_{YX} C_{XX}^{-1}(X - \mu_X)\right)$$

- Using the Matrix Inversion Identity

For any jointly Gaussian RV

$$P(X,Y) = Const \exp\left(-0.5(Z - \mu_Z)^T C_Z^{-1}(Z - \mu_Z)\right) =$$

$$= const \exp(-0.5 Quadratic(X) +$$

$$-0.5\left(Y - \mu_Y - C_{YX} C_{XX}^{-1}(X - \mu_X)\right)^T \left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} \left(Y - \mu_Y - C_{YX} C_{XX}^{-1}(X - \mu_X)\right))$$

$$P(Y \mid X) =$$

$$K \exp\left(-0.5\left(Y - \mu_Y - C_{YX} C_{XX}^{-1}(X - \mu_X)\right)^T \left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)^{-1} \left(Y - \mu_Y - C_{YX} C_{XX}^{-1}(X - \mu_X)\right)\right)$$

$$= Gaussian\left(Y; \left(Y - \mu_Y - C_{YX} C_{XX}^{-1}(X - \mu_X)\right), \left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)\right)$$

- The conditional of Y is a Gaussian

Estimating P(s | o)

Dropping subscript t and $o_{0:t-1}$ for brevity

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

Assuming γ is 0 mean

$$o = Bs + \gamma \qquad P(\gamma) = \frac{1}{\sqrt{(2\pi)^d \mid \Theta_\gamma \mid}} \exp\left(-0.5\varepsilon^T \Theta_\gamma^{-1} \varepsilon\right)$$

- Consider the joint distribution of o and s

$$O = \begin{bmatrix} o \\ s \end{bmatrix}$$

- O is a linear function of s
  - Hence O is also Gaussian

$$P(O) = Gaussian(O; \mu_O, \Theta_O)$$

The probability distribution of O

$$o = Bs + \gamma \qquad O = \begin{bmatrix} o \\ s \end{bmatrix}$$

$$P(s) = Gaussian(s; \bar{s}, R) \qquad P(\gamma) = Gaussian(\gamma; 0, \Theta_\gamma)$$

$$P(O) = Gaussian(O; \mu_O, \Theta_O)$$

$$\mu_O = E[O] = E[\begin{bmatrix} o \\ s \end{bmatrix}] = \begin{bmatrix} E[o] \\ E[s] \end{bmatrix} = \begin{bmatrix} B\bar{s} \\ \bar{s} \end{bmatrix}$$

$$\mu_O = \begin{bmatrix} B\bar{s} \\ \bar{s} \end{bmatrix}$$

The probability distribution of O

$$P(O) = Gaussian(O; \mu_O, \Theta_O) \qquad \mu_O = \begin{bmatrix} B\bar{s} \\ \bar{s} \end{bmatrix} \qquad o = Bs + \gamma$$

$$P(\gamma) = Gaussian(\gamma; 0, \Theta_\gamma) \qquad P(s) = Gaussian(s; \bar{s}, R)$$

$$\Theta_O = E[(O-\mu_O)(O-\mu_O)^T] = E\left[\begin{bmatrix} o - B\bar{s} \\ s - \bar{s} \end{bmatrix}\begin{bmatrix} (o-B\bar{s})^T & s^T - \bar{s}^T \end{bmatrix}\right]$$

$$\Theta_O = E[(O-\mu_O)(O-\mu_O)^T] = E\left[\begin{bmatrix} B(s-\bar{s})+\gamma \\ s - \bar{s} \end{bmatrix}\begin{bmatrix} (s-\bar{s})^T B^T + \gamma^T & (s-\bar{s})^T \end{bmatrix}\right]$$

$$\Theta_O = E\begin{bmatrix} (B(s-\bar{s})+\gamma)((s-\bar{s})^T B^T + \gamma^T) & (B(s-\bar{s})+\gamma)(s-\bar{s})^T \\ (s-\bar{s})(B(s-\bar{s})+\gamma)^T & (s-\bar{s})(s-\bar{s})^T \end{bmatrix}$$

The probability distribution of O

$$P(O) = Gaussian(O; \mu_O, \Theta_O) \qquad \mu_O = \begin{bmatrix} B\bar{s} \\ \bar{s} \end{bmatrix} \qquad o = Bs + \gamma$$

$$P(\gamma) = Gaussian(\gamma; 0, \Theta_\gamma) \qquad P(s) = Gaussian(s; \bar{s}, R)$$

$$\Theta_O = E\begin{bmatrix} (B(s-\bar{s})+\gamma)((s-\bar{s})^T B^T + \gamma^T) & (B(s-\bar{s})+\gamma)(s-\bar{s})^T \\ (s-\bar{s})(B(s-\bar{s})+\gamma)^T & (s-\bar{s})(s-\bar{s})^T \end{bmatrix}$$

$$\Theta_O = \begin{bmatrix} E[(B(s-\bar{s})+\gamma)((s-\bar{s})^T B^T + \gamma^T)] & E[(B(s-\bar{s})+\gamma)(s-\bar{s})^T] \\ E[(s-\bar{s})(B(s-\bar{s})+\gamma)^T] & E[(s-\bar{s})(s-\bar{s})^T] \end{bmatrix}$$

$$\Theta_O = \begin{bmatrix} BRB^T + \Theta_\gamma & BR \\ RB^T & R \end{bmatrix}$$

The probability distribution of O

$$o = Bs + \gamma$$

$$P(\gamma) = Gaussian(\gamma; 0, \Theta_\gamma) \qquad P(s) = Gaussian(s; \bar{s}, R)$$

$$O = \begin{bmatrix} o \\ s \end{bmatrix} \qquad P(O) = Gaussian(O; \mu_O, \Theta_O)$$

$$\Theta_O = \begin{bmatrix} BRB^T + \Theta_\gamma & BR \\ RB^T & R \end{bmatrix} \qquad \mu_O = \begin{bmatrix} B\bar{s} \\ \bar{s} \end{bmatrix}$$

The probability distribution of O

$$P(O \mid o_{0:t-1}) = P(o, s \mid o_{0:t-1}) = Gaussian(O; \mu_O, \Theta_O)$$

$$C\exp\left(-0.5\begin{bmatrix} (o - B\bar{s}) & (s-\bar{s}) \end{bmatrix}^T \begin{bmatrix} BRB^T + \Theta_\gamma & BR \\ RB^T & R \end{bmatrix}^{-1} \begin{bmatrix} o - B\bar{s} \\ s - \bar{s} \end{bmatrix}\right)$$

- Writing it out in extended form

Recall: For any jointly Gaussian RV

$$P(X,Y) = Const\exp\left(-0.5(Z-\mu_Z)^T C_Z^{-1}(Z-\mu_Z)\right) =$$

$$P(Y \mid X) =$$

$$= Gaussian\left(Y; \left(Y - \mu_Y - C_{YX}C_{XX}^{-1}(X-\mu_X)\right), \left(C_{YY} - C_{XY}^T C_{XX}^{-1} C_{XY}\right)\right)$$

- Applying it to:

$$P(O \mid o_{0:t-1}) = P(o, s \mid o_{0:t-1}) = Gaussian(O; \mu_O, \Theta_O)$$

$$C\exp\left(-0.5\begin{bmatrix} (o - B\bar{s}) & (s-\bar{s}) \end{bmatrix}^T \begin{bmatrix} BRB^T + \Theta_\gamma & BR \\ RB^T & R \end{bmatrix}^{-1} \begin{bmatrix} o - B\bar{s} \\ s - \bar{s} \end{bmatrix}\right)$$

Stable Estimation

$$P(O \mid o_{0:t-1}) = P(o, s \mid o_{0:t-1}) = Gaussian(O; \mu_Y, \Theta_O)$$

$$C\exp\left(-0.5\begin{bmatrix} (o - B\bar{s}) & (s-\bar{s}) \end{bmatrix}^T \begin{bmatrix} BRB^T + \Theta_\gamma & BR \\ RB^T & R \end{bmatrix}^{-1} \begin{bmatrix} o - B\bar{s} \\ s - \bar{s} \end{bmatrix}\right)$$

- The conditional distribution of s

$$P(s \mid o_{0:t}) = Gaussian\left(s; (I - RB^T(BRB^T + \Theta_\gamma)^{-1}B)\bar{s} + RB^T(BRB^T + \Theta_\gamma)^{-1}o, \left(R - RB^T(BRB^T + \Theta_\gamma)^{-1}BR\right)\right)$$

- Note that we are not computing Θ_γ^{-1} in this formulation

The Kalman filter

- The actual state estimate is the *mean* of the updated distribution
- Predicted state at time *t*

$$\bar{s}_t = s_t^{pred} = mean[P(s_t \mid o_{0:t-1})] = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

- Updated estimate of state at time *t*

$$P(s_t \mid o_{0:t}) = Gaussian\left(s; (I - RB^T(BRB^T + \Theta_\gamma)^{-1}B)\bar{s} + RB^T(BRB^T + \Theta_\gamma)^{-1}o, \left(R - RB^T(BRB^T + \Theta_\gamma)^{-1}BR\right)\right)$$

$$\hat{s}_t = mean[P(s_t \mid o_{0:t})] = (I - R_t B_t^T (B_t R_t B_t^T + \Theta_\gamma)^{-1} B_t)\bar{s}_t + R_t B_t^T (B_t R_t B_t^T + \Theta_\gamma)^{-1} o_t$$

The Kalman filter

- Prediction

$$\bar{s}_t = s_t^{pred} = mean[P(s_t \mid o_{0:t-1})] = A_t \hat{s}_{t-1} + \mu_\varepsilon$$

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

- Update

$$\hat{s}_t = \left(I - R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma\right)^{-1} B_t\right)\bar{s}_t + R_t B_t^T \left(B_t R_t B_t^T + \Theta_\gamma\right)^{-1} o_t$$

$$\hat{R}_t = R_t - R_t B_t^T (B_t R_t B_t^T + \Theta_\gamma)^{-1} B_t R_t$$

The Kalman filter

- 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

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

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

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

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

Solutions

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

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

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