Machine Learning for Signal Processing
Predicting and Estimation from Time Series

Bhiksha Raj
Class 22. 14 Nov 2013
Administrivia

• No class on Tuesday..

• Project Demos: 5\textsuperscript{th} December (Thursday).
  – Before exams week
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!

- 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

<table>
<thead>
<tr>
<th>State</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idling</td>
<td>0.25</td>
</tr>
<tr>
<td>Accelerating</td>
<td>0.25</td>
</tr>
<tr>
<td>Cruising</td>
<td>0.25</td>
</tr>
<tr>
<td>Decelerating</td>
<td>0.25</td>
</tr>
</tbody>
</table>
The first observation

At T=0 we observe the sound level $x_0 = 67\text{dB SPL}$

- The observation modifies our belief in the state of the system

- $P(x_0 | \text{idle}) = 0$
- $P(x_0 | \text{deceleration}) = 0.0001$
- $P(x_0 | \text{acceleration}) = 0.7$
- $P(x_0 | \text{cruising}) = 0.5$
  - 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 $x_0$

- $P(\text{state} \mid x_0) = C \ P(\text{state})P(x_0 \mid \text{state})$
  - $P(\text{idle} \mid x_0) = 0$
  - $P(\text{deceleration} \mid x_0) = C \ 0.000025$
  - $P(\text{cruising} \mid x_0) = C \ 0.125$
  - $P(\text{acceleration} \mid x_0) = C \ 0.175$

- Normalizing
  - $P(\text{idle} \mid x_0) = 0$
  - $P(\text{deceleration} \mid x_0) = 0.000083$
  - $P(\text{cruising} \mid x_0) = 0.42$
  - $P(\text{acceleration} \mid x_0) = 0.57$
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 at $T=1$

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

- In general, for any state $S$
  - $P(S_{T=1} \mid x_0) = \sum_{S_{T=0}} P(S_{T=0} \mid x_0) P(S_{T=1} \mid S_{T=0})$
Predicting the state at $T = 1$

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

<table>
<thead>
<tr>
<th>State</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idling</td>
<td>0.0</td>
</tr>
<tr>
<td>Accelerating</td>
<td>0.57</td>
</tr>
<tr>
<td>Cruising</td>
<td>0.42</td>
</tr>
<tr>
<td>Decelerating</td>
<td>8.3 x 10^{-5}</td>
</tr>
</tbody>
</table>

- Idling: 0.0
- Accelerating: 0.57
- Cruising: 0.42
- Decelerating: 8.3 x 10^{-5}

$P(S_{T=1} \mid x_0)$

- 0.33
- 0.33
- 0.33

$2.1 \times 10^{-5}$
Updating after the observation at T=1

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

<table>
<thead>
<tr>
<th>State</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idling</td>
<td>0.0</td>
</tr>
<tr>
<td>Accelerating</td>
<td>0.57</td>
</tr>
<tr>
<td>Cruising</td>
<td>0.42</td>
</tr>
<tr>
<td>Decelerating</td>
<td>0.713</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idling</td>
<td>0.0014</td>
</tr>
<tr>
<td>Accelerating</td>
<td>0.0</td>
</tr>
<tr>
<td>Cruising</td>
<td>0.713</td>
</tr>
<tr>
<td>Decelerating</td>
<td>0.285</td>
</tr>
</tbody>
</table>

14 Nov 2013 11-755/18797
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:
  \[ P(x_{0:T}, S_T) = P(x_T | S_T) \sum_{S_{T-1}} P(S_{T-1} | x_{0:T-1}) P(S_T | S_{T-1}) \]

• Normalized:
  \[ P(S_T | x_{0:T}) = \left( \sum_{S'} P(x_{0:T}, S') \right)^{-1} P(x_{0:T}, S_T) = C \cdot P(x_{0:T}, S_T) \]
Decomposing the forward algorithm

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

• Predict:
  - $P(x_{0:T-1}, S_T) = \sum_{S_{T-1}} P(x_{0:T-1}, S_{T-1}) P(S_T|S_{T-1})$

• Update:
  - $P(x_{0:T}, S_T) = P(x_T|S_T) P(x_{0:T-1}, S_T)$
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:
  \[ P(x_T|x_{0:T-1}) = \sum_{S_{T-1}} P(S_{T-1}|x_{0:T-1}) P(S_T|x_{0:T-1}) \]
- The actual observation can be predicted from \( P(x_T|x_{0:T-1}) \)
Predicting the next observation

- **MAP estimate:**
  - \( \arg\max_{x_T} P(x_T|x_{0:T-1}) \)

- **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, cruising 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 \( t \)
  – \( \varepsilon_t \) 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 \( t \)
  – \( \gamma_t \) 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

- 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)

\[
\begin{align*}
s_t &= f(s_{t-1}, \varepsilon_t) \\
o_t &= g(s_t, \gamma_t)
\end{align*}
\]
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

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

- $P(s_0 | o_0) = C \cdot P_0(s_0)P(o_0 | s_0)$
The observation probability: $P(o \mid s)$

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

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

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

• \( P(o|s) = ? \)

\[ o_t = g(s_t, \gamma_t) \]

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

• 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 $t=1$

$$P(s_1 | o_0) = \int P(s_1, s_0 | o_0) ds_0 = \int P(s_1 | s_0) P(s_0 | o_0) ds_0$$

- State progression function:

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

  - $\varepsilon_t$ 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

Prediction at time 0:
\[ P(s_0) = \pi(s_0) \]

Update after \( O_0 \):
\[ P(s_0 | O_0) = C \pi(s_0) P(O_0 | s_0) \]

Prediction at time 1:
\[ P(s_1 | O_0) = \sum_{s_0} P(s_0 | O_0) P(s_1 | s_0) \]

Update after \( O_1 \):
\[ P(s_1 | O_0, O_1) = C P(s_1 | O_0) P(O_1 | s_1) \]

\[ s_t = f(s_{t-1}, \varepsilon_t) \]
\[ o_t = g(s_t, \gamma_t) \]
Discrete vs. Continuous State Systems

Prediction at time $t$

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

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

Update after $O_t$:

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

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

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

$o_t = g(s_t, \gamma_t)$
Discrete vs. Continuous State Systems

Parameters

Initial state prob. \( \pi \)

Transition prob \( \{T_{ij}\} = P(s_t = j \mid s_{t-1} = i) \)

Observation prob \( P(O \mid s) \)

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

Observation function \( o_t = g(s_t, \gamma_t) \)

\[ P(s) \]

\[ P(s_t \mid s_{t-1}) \]

\[ P(o \mid s) \]
Special case: Linear Gaussian model

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

\[ o_t = B_t s_t + \gamma_t \]

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

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

- A **linear** state dynamics equation
  - Probability of state driving term \( \varepsilon \) is Gaussian
  - Sometimes viewed as a driving term \( \mu_\varepsilon \) and additive zero-mean noise
- A **linear** observation equation
  - Probability of observation noise \( \gamma \) is Gaussian
- \( A_t, B_t \) and Gaussian parameters assumed known
  - May vary with time
The initial state probability

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

\[P_0(s) = \text{Gaussian}(s; \bar{s}, R)\]

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

\[s_t = A_t s_{t-1} + \epsilon_t\]

\[o_t = B_t s_t + \gamma_t\]
The observation probability

\[ o_t = B_t s_t + \gamma_t \]

\[ P(\gamma) = \text{Gaussian}(\gamma; \mu_{\gamma}, \Theta_{\gamma}) \]

\[ P(o_t \mid s_t) = \text{Gaussian}(o_t; \mu_{\gamma} + B_t s_t, \Theta_{\gamma}) \]

- 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 $T=0$

\[ P(s_0|o_0) = C \, P(s_0) \, P(o_0|s_0) \]

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

$P(o_0|s_0) = \text{Gaussian}(o_0; \mu_\gamma + B_0 s_0, \Theta_\gamma)$

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

\[ o_t = B_t s_t + \gamma_t \]

\[ P(\gamma) = \mathcal{N}(\gamma; \mu_\gamma, \Theta_\gamma) \]
Note 1: product of two Gaussians

• The product of two Gaussians is a Gaussian

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

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

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

Not a good estimate --
The updated state probability at $T=0$

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

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

$P(o_0|s_0) = \text{Gaussian}(o_0; \mu_\gamma + B_0 s_0, \Theta_\gamma)$

$P(s_0|o_0) = \text{Gaussian}(s_0; \hat{s}_0, \hat{R}_0)$
The state transition probability

\[ s_t = A_t s_{t-1} + \epsilon_t \]

\[ P(\epsilon) = \text{Gaussian}(\epsilon; \mu_\epsilon, \Theta_\epsilon) \]

\[ P(s_t \mid s_{t-1}) = \text{Gaussian}(s_t; \mu_\epsilon + A_t s_{t-1}, \Theta_\epsilon) \]

- 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

\[
\int_{-\infty}^{\infty} \text{Gaussian}(x; \mu_x, \Theta_x) \text{Gaussian}(y; Ax + b, \Theta_y) \, dx = \\
\int_{-\infty}^{\infty} C_1 \exp\left(-0.5(x - \mu_x)^T \Theta_x^{-1} (x - \mu_x)\right) C_2 \exp\left(-0.5(y - Ax - b)^T \Theta_y^{-1} (y - Ax - b)\right) \, dx \\
= \text{Gaussian}\left(y; A\mu_x + b, \Theta_y + A\Theta_x A^T\right)
\]
Note 2: integral of product of two Gaussians

\[ y = Ax + e \quad x \sim N(\mu_x, \Theta_x) \quad e \sim N(b, \Theta_y) \]

\[ P(y) = N(A\mu_x + b, \Theta_y + A\Theta_x A^T) \]

- \( P(y) \) is the integral of the product of two Gaussians is a Gaussian

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

\[ = \text{Gaussian}(y; A\mu_x + b, \Theta_y + A\Theta_x A^T) \]
The predicted state probability at $t=1$

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

\[
P(s_1 \mid s_0) = \text{Gaussian}(s_1; \mu_\varepsilon + A_1 s_0, \Theta_\varepsilon)
\]

\[
P(s_0 \mid o_0) = \text{Gaussian}(s_0; \hat{s}_0, \hat{R}_0)
\]

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

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

- Remains Gaussian

\[
s_t = A_t s_{t-1} + \varepsilon_t
\]
The updated state probability at $T=1$

\[ P(s_1 \mid o_{0:1}) = C \cdot P(s_1 \mid o_0) \cdot P(o_1 \mid s_1) \]

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

\[ P(o_1 \mid s_1) = \text{Gaussian}(o_1; \mu_\gamma + B_1 s_1, \Theta_\gamma) \]

\[ P(s_1 \mid o_{0:1}) = \text{Gaussian}(s_1; \hat{s}_1, \hat{R}_1) \]
The Kalman Filter!

• Prediction at $T$

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

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

• Update at $T$

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

$$o_t = B_t s_t + \gamma_t$$

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

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

\[ s_t = A_t s_{t-1} + \epsilon_t \]

\[ o_t = B_t s_t + \gamma_t \]

\[ P(s) = P(s_0) = P(s) \]

\[ P(s_0 \mid O_0) = C \ P(s_0) \ P(O_0 \mid s_0) \]

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

\[ P(s_1 \mid O_{0:1}) = C \ P(s_1 \mid O_0) \ P(O_1 \mid s_0) \]

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

\[ P(s_2 \mid O_{0:2}) = C \ P(s_2 \mid O_{0:1}) \ P(O_2 \mid s_2) \]

All distributions remain Gaussian
The Kalman filter

- The actual state estimate is the \textit{mean} of the updated distribution

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

- Predicted state at time \( t \)

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

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

• 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
Conditional Probability of $y \mid x$

- If $P(x,y)$ is Gaussian:
  $$P(y, x, k) = N\left( \begin{bmatrix} \mu_{k,x} \\ \mu_{k,y} \end{bmatrix}, \begin{bmatrix} C_{k,xx} & C_{k,xy} \\ C_{k,yx} & C_{k,yy} \end{bmatrix} \right)$$

- 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
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

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

\[ C_Z^{-1} = \begin{bmatrix} C_{xx}^{-1} + C_{xx}^{-1} C_{xy} \left( C_{yy} - C_{xy}^T C_{xx}^{-1} C_{xy} \right)^{-1} C_{xy}^T C_{xx}^{-1} & -C_{xx}^{-1} C_{xy} \left( C - C_{xy}^T C_{xx}^{-1} C_{xy} \right)^{-1} \\ -\left( C_{yy} - C_{xy}^T C_{xx}^{-1} C_{xy} \right)^{-1} C_{xy}^T C_{xx}^{-1} & \left( C_{yy} - C_{xy}^T C_{xx}^{-1} C_{xy} \right)^{-1} \end{bmatrix} \]

\[
(Z - \mu_Z)^T C_Z^{-1} (Z - \mu_Z) = \text{Quadratic}(X) + \\
(Y - \mu_y - C_{yx} C_{xx}^{-1} (X - \mu_x))^T \left( C_{yy} - C_{xy}^T C_{xx}^{-1} C_{xy} \right)^{-1} (Y - \mu_y - C_{yx} C_{xx}^{-1} (X - \mu_x))
\]

- Using the Matrix Inversion Identity
For any jointly Gaussian RV

\[ P(X, Y) = \text{Const} \exp \left( -0.5 (Z - \mu_z)^T C_z^{-1} (Z - \mu_z) \right) = \]

\[ = \text{const} \exp ( -0.5 \text{Quadratic}(X) + \]

\[ -0.5 (Y - \mu_y - C_{yx} C_{xx}^{-1} (X - \mu_x))^T (C_{yy} - C_{xy}^T C_{xx}^{-1} C_{xy})^{-1} (Y - \mu_y - C_{yx} C_{xx}^{-1} (X - \mu_x)) \]

\[ P(Y \mid X) = \]

\[ K \exp \left( -0.5 (Y - \mu_y - C_{yx} C_{xx}^{-1} (X - \mu_x))^T (C_{yy} - C_{xy}^T C_{xx}^{-1} C_{xy})^{-1} (Y - \mu_y - C_{yx} C_{xx}^{-1} (X - \mu_x)) \right) \]

\[ = \text{Gaussian}(Y; \mu_y + C_{yx} C_{xx}^{-1} (X - \mu_x), (C_{yy} - C_{xy}^T C_{xx}^{-1} C_{xy}) \]

• The conditional of Y is a Gaussian
Conditional Probability of $y \mid x$

- If $P(x,y)$ is Gaussian:
  \[
P(y,x,k) = N\left(\begin{bmatrix} \mu_{k,x} \\ \mu_{k,y} \end{bmatrix}, \begin{bmatrix} C_{k,xx} & C_{k,xy} \\ C_{k,yx} & C_{k,yy} \end{bmatrix}\right)
  \]

- 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
Estimating $P(s | o)$

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

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

Assuming $\gamma$ is 0 mean

$$o = Bs + \gamma$$

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

- Consider the joint distribution of $o$ and $s$

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

$$P(O) = Gaussian(O; \mu_O, \Theta_O)$$
The joint PDF of $o$ and $s$

- $o = Bs + \gamma$
- $\mu_o = B\bar{s}$
- $C_{o,o} = BRB^T + \Theta_\gamma$
- $P(o \mid o_{0:t-1}) = \text{Gaussian}(B\bar{s}, BRB^T + \Theta_\gamma)$
- $P(s \mid o_{0:t-1}) = \text{Gaussian}(s; \bar{s}, R)$
- $P(\gamma) = \text{Gaussian}(0, \Theta_\gamma)$

- $o$ is Gaussian. Its cross covariance with $s$: $C_{o,s} = BR$
The probability distribution of $O$

$$o = Bs + \gamma$$

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

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

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

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

$$\mu_O = E[O] = E\left[ \begin{pmatrix} o \\ s \end{pmatrix} \right] = \begin{pmatrix} E[o] \\ E[s] \end{pmatrix} = \begin{pmatrix} Bs \bar{s} \\ \bar{s} \end{pmatrix}$$

$$\mu_O = \begin{pmatrix} Bs \bar{s} \\ \bar{s} \end{pmatrix}$$
The probability distribution of $O$

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

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

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

$$\Theta_O = \begin{bmatrix} C_{o,o} & C_{o,s} \\ C_{s,o} & C_{s,s} \end{bmatrix}$$

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

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

$$o = Bs + \gamma$$

$$C_{o,o} = BRB^T + \Theta_\gamma$$

$$C_{o,s} = BR$$

$$\Theta_O = \begin{bmatrix} BRB^T + \Theta_\gamma & BR \\ RB^T & R \end{bmatrix}$$
The probability distribution of $O$

$$ o = Bs + \gamma $$

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

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

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

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

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

$$ \mu_O = \begin{bmatrix} B\bar{s} \\ \bar{s} \end{bmatrix} $$
The probability distribution of $O$

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

- Writing it out in extended form

$$C \exp\left(-0.5\left[(o - B\bar{s}) (s - \bar{s})\right]^T \left[\begin{array}{cc} BRB^T + \Theta & BR \\ RB^T & R \end{array}\right]^{-1} \left[\begin{array}{c} o - B\bar{s} \\ s - \bar{s} \end{array}\right]\right)$$
Recall: For any jointly Gaussian RV

\[ P(Y \mid X) = \text{Gaussian}(Y; \mu_Y + C_{YX} C_{XX}^{-1} (X - \mu_X), (C_{YY} - C_{XY} C_{XX}^{-1} C_{XY})) \]

- Applying it to our problem (replace Y by s, X by o):

\[
P(s \mid o_{0:t}) = \text{Gaussian}(s; \mu, \Theta)
\]

\[
\mu = (I - RB^T (BRB^T + \Theta \gamma)^{-1} B) \bar{s} + RB^T (BRB^T + \Theta \gamma)^{-1} o
\]

\[
\Theta = R - RB^T (BRB^T + \Theta \gamma)^{-1} BR
\]
Stable Estimation

\[ P(s \mid o_{0:t}) = Gaussian(s; \mu_{s\mid o_{1:t}}, \Theta_{s\mid o_{1:t}}) \]

\[
\mu_{s\mid o_{1:t}} = (I - RB^T (BRB^T + \Theta_{\gamma})^{-1} B)\bar{s} + RB^T (BRB^T + \Theta_{\gamma})^{-1} o_t
\]

\[
\Theta_{s\mid o_{1:t}} = R - RB^T (BRB^T + \Theta_{\gamma})^{-1} BR
\]

- Note that we are not computing \( \Theta_{\gamma}^{-1} \) in this formulation
The Kalman filter

• The actual state estimate is the *mean* of the updated distribution

\[
s_t = A_t s_{t-1} + \epsilon_t
\]

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

• Predicted state at time \( t \)

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

• Updated estimate of state at time \( t \)

\[
o_t = B_t s_t + \gamma_t
\]
The Kalman filter

• Prediction

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

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

• Update

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

$$\hat{R}_t = R_t - R_t B_t^T \left( B_t R_t B_t^T + \Theta_\gamma \right)^{-1} B_t R_t$$
The Kalman filter

• Prediction

\[ \bar{s}_t = A_t \hat{s}_{t-1} + \mu_{\epsilon} \]

\[ s_t = A_t s_{t-1} + \epsilon_t \]

\[ R_t = \Theta_{\epsilon} + A_t \hat{R}_{t-1} A_t^T \]

• Update

\[ K_t = R_t B_t^T (B_t R_t B_t^T + \Theta_\gamma)^{-1} \]

\[ \hat{s}_t = \bar{s}_t + K_t (o_t - B_t \bar{s}_t) \]

\[ \hat{R}_t = (I - K_t B_t) R_t \]
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} + \epsilon_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: $s_t = A_t s_{t-1} + G_t u_t + \epsilon_t$
  – 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

- \( \varepsilon \) and/or \( \gamma \) 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

- \( \varepsilon \) and/or \( \gamma \) may not be Gaussian
  - Gaussian based update rules no longer valid
  - Particle Filters