Maximum Likelihood (ML), Expectation Maximization (EM)

Pieter Abbeel
UC Berkeley EECS

Many slides adapted from Thrun, Burgard and Fox, Probabilistic Robotics

Outline

- Maximum likelihood (ML)
- Priors, and maximum a posteriori (MAP)
- Cross-validation
- Expectation Maximization (EM)

Thumbtack

- Let θ = P(up), 1-θ = P(down)
- How to determine θ ?

- Empirical estimate: 8 up, 2 down → $\theta = \frac{8}{2+8} = 0.8$

Date Time

LESSON 2·6

A Thumbtack Experiment

Make a guess: If you drop a thumbtack, is it more likely to land with the point up or with the point down? point down

The experiment described below will enable you to make an estimate of the chance that a thumbtack will land point down.

1. Work with a partner. You should have 10 thumbtacks and 1 small cup. Do the experiment at your desk or a table so you are working over a smooth, hard surface.

   Place the 10 thumbtacks inside the cup. Shake the cup a few times, and then carefully drop the tacks onto the desk surface. Record the number of thumbtacks that land point up and the number that land point down.

   Toss the 10 thumbtacks 9 more times and record the results each time.

Toss	Number Landing Point Up	Number Landing Point Down
1	8	2
2	6	4
3	6	4
4	6	4
5	8	2
6	9	1
7	8	2
8	6	4
9	10	0
10	10	0
	Total Up = 77	Total Down = 23

2. In making your 10 tosses, you dropped a total of 100 thumbtacks.

   What fraction of the thumbtacks landed point down? $\frac{23}{100}$

3. Write this fraction on a small stick-on note. Also write it as a decimal and as a percent.

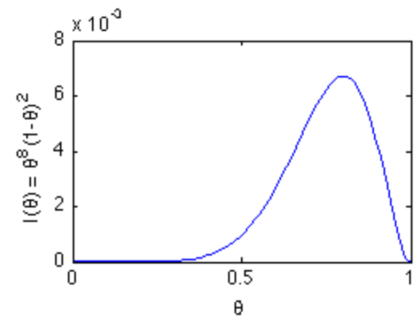
4. For the whole class, the chance that a tack will land point down is Unlikely.

- http://web.me.com/todd6ton/Site/Classroom_Blog/Entries/2009/10/7_A_Thumbtack_Experiment.html

Maximum Likelihood

- θ = P(up), 1-θ = P(down)
- Observe:

- Likelihood of the observation sequence depends on θ:



$$\begin{aligned} l(\theta) &= \theta(1-\theta)\theta(1-\theta)\theta\theta\theta\theta\theta\theta\theta\theta \\ &= \theta^8(1-\theta)^2 \end{aligned}$$

- Maximum likelihood finds

$$\arg\max_\theta l(\theta) = \arg\max_\theta \theta^8(1-\theta)^2$$

$$\frac{\partial}{\partial\theta} l(\theta) = 8\theta^7(1-\theta)^2 - 2\theta^8(1-\theta) = \theta^7(1-\theta)(8(1-\theta)-2\theta) = \theta^7(1-\theta)(8-10\theta)$$

→ extrema at θ = 0, θ = 1, θ = 0.8

→ Inspection of each extremum yields $\theta_{ML} = 0.8$

Maximum Likelihood

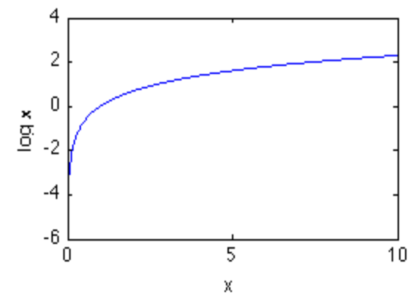
- More generally, consider binary-valued random variable with θ = P(1), 1-θ = P(0), assume we observe n_1 ones, and n_0 zeros
  - Likelihood: $l(\theta) = \theta^{n_1}(1-\theta)^{n_0}$
  - Derivative: $$\begin{aligned} \frac{\partial}{\partial\theta} l(\theta) &= n_1\theta^{n_1-1}(1-\theta)^{n_0} - n_0\theta^{n_1}(1-\theta)^{n_0-1} \\ &= \theta^{n_1-1}(1-\theta)^{n_0-1}\left(n_1(1-\theta) - n_0\theta\right) \\ &= \theta^{n_1-1}(1-\theta)^{n_0-1}(n_1 - (n_1+n_0)\theta) \end{aligned}$$
  - Hence we have for the extrema:

$$\theta = 0, \quad \theta = 1, \quad \theta = \frac{n_1}{n_0+n_1}$$

    - n1/(n0+n1) is the maximum
- = empirical counts.

Log-likelihood

- The function $\log : \mathbb{R}^+ \rightarrow \mathbb{R} : x \rightarrow \log(x)$



is a monotonically increasing function of x

- Hence for any (positive-valued) function f:

$$\arg\max_\theta f(\theta) = \arg\max_\theta \log f(\theta)$$

- Often more convenient to optimize log-likelihood rather than likelihood

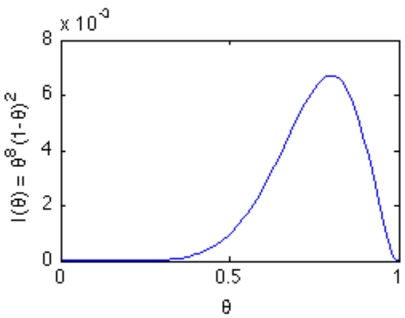
- Example:

$$\begin{aligned} \log l(\theta) &= \log \theta^{n_1}(1-\theta)^{n_0} \\ &= n_1 \log\theta + n_0 \log(1-\theta) \end{aligned}$$

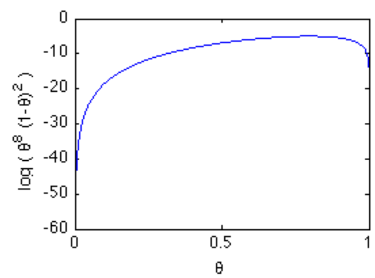
$$\begin{aligned} \frac{\partial}{\partial\theta}\log l(\theta) &= n_1 \frac{1}{\theta} + n_0 \frac{-1}{1-\theta} = \frac{n_1 - (n_1+n_0)\theta}{\theta(1-\theta)} \\ &\rightarrow \quad \theta = \frac{n_1}{n_1+n_0} \end{aligned}$$

Log-likelihood ←→ Likelihood

- Reconsider thumbtacks: 8 up, 2 down
  - Likelihood
  - Log-likelihood



Not Concave



Concave

- Definition: A function f is concave if and only

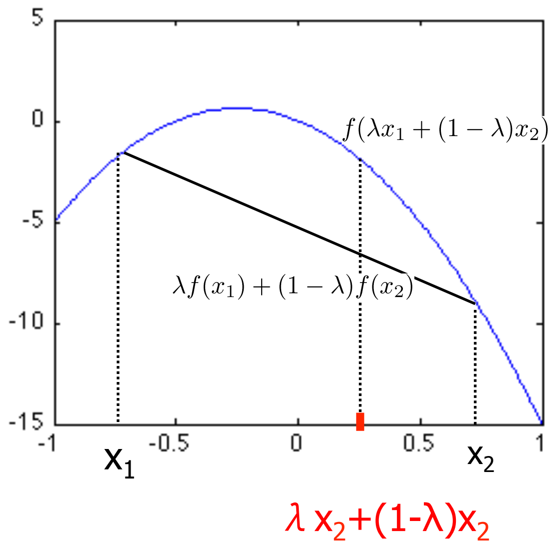
$$\forall x_1, x_2, \quad \forall \lambda \in (0,1), f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda) f(x_2)$$

- Concave functions are generally easier to maximize then non-concave functions

Concavity and Convexity

f is ***concave*** if and only

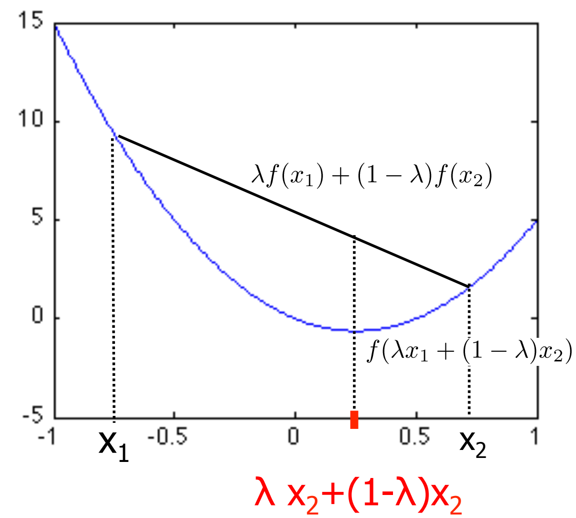
$$\forall x_1, x_2, \quad \forall \lambda \in (0,1),$$
$$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda) f(x_2)$$



"Easy" to maximize

f is ***convex*** if and only

$$\forall x_1, x_2, \quad \forall \lambda \in (0,1),$$
$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$



"Easy" to minimize

ML for Multinomial

$p(x = k; \theta) = \theta_k$

- Consider having received samples $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

$$\begin{aligned}
\log l(\theta) &= \log \prod_{i=1}^{m} \theta_1^{1\{x^{(i)}=1\}} \theta_2^{1\{x^{(i)}=2\}} \cdots \theta_{K-1}^{1\{x^{(i)}=K-1\}} (1 - \theta_1 - \theta_2 - \ldots - \theta_{K-1})^{1\{x^{(i)}=K\}} \\
&= \sum_{i=1}^{m} 1\{x^{(i)} = 1\} \log \theta_1 + 1\{x^{(i)} = 2\} \log \theta_2 + \cdots + 1\{x^{(i)} = K-1\} \log \theta_{K-1} + 1\{x^{(i)} = K\} \log(1 - \theta_1 - \theta_2 - \cdots - \theta_{K-1}) \\
&= \sum_{k=1}^{K-1} n_k \log \theta_k + n_K \log(1 - \theta_1 - \theta_2 - \cdots - \theta_{K-1})
\end{aligned}$$

$$\frac{\partial}{\partial \theta_k} \log l(\theta) = \frac{n_k}{\theta_k} - n_K \frac{1}{1 - \theta_1 - \theta_2 - \cdots - \theta_{K-1}}$$

$$\rightarrow \theta_k^{\mathrm{ML}} = \frac{n_k}{\sum_{j=1}^{K} n_j}$$

ML for Fully Observed HMM

- Given samples $\{x_0, z_0, x_1, z_1, x_2, z_2, \ldots, x_T, z_T\},\ x_t \in \{1, 2, \ldots, I\}, z_t \in \{1, 2, \ldots, K\}$
- Dynamics model: $P(x_{t+1} = i | x_t = j) = \theta_{i|j}$
- Observation model: $P(z_t = k | z_t = l) = \gamma_{k|l}$

$$
\begin{aligned}
\log l(\theta, \gamma) &= \log P(x_0) \prod_{t=1}^{T} P(x_t | x_{t-1}; \theta) P(z_t | x_t; \gamma) \\
&= \log P(x_0) \sum_{t=1}^{T} \log \theta_{x_t | x_{t-1}} + \sum_{t=1}^{T} \log \gamma_{z_t | x_t} \\
&= \log P(x_0) \sum_{i=1}^{I} \sum_{j=1}^{I} \log \theta_{i|j}^{n_{(i,j)}} + \sum_{k=1}^{K} \sum_{l=1}^{K} \log \gamma_{k|l}^{m_{(k,l)}}
\end{aligned}
$$

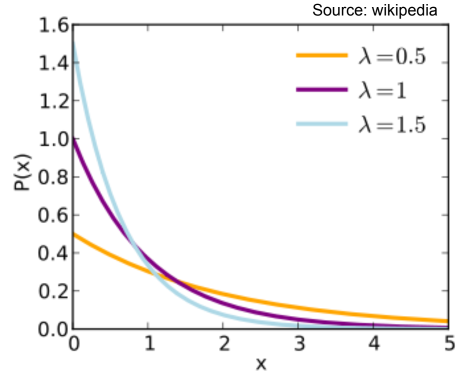
$n_{(i,j)}$: number of occurences of $x_t = i, x_{t+1} = j$.
$m_{(k,l)}$: number of occurences of $x_t = k, z_t = l$.

→ Independent ML problems for each $\theta_{\cdot|j}$ and each $\gamma_{\cdot|l}$

$$
\theta_{i|j} = \frac{n_{(i,j)}}{\sum_{i'=1}^{I} n_{(i',j)}} \qquad \gamma_{k|l} = \frac{m_{(k,l)}}{\sum_{k'=1}^{K} m_{(k',l)}}
$$

ML for Exponential Distribution

Source: wikipedia

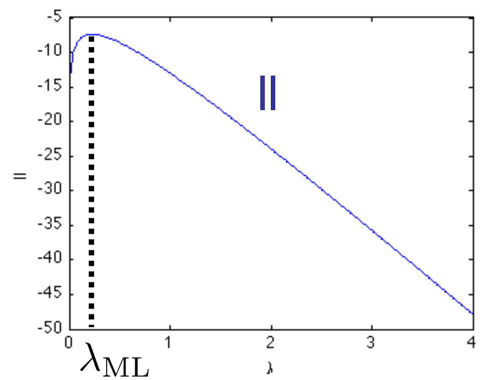


$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

- Consider having received samples
  - 3.1, 8.2, 1.7

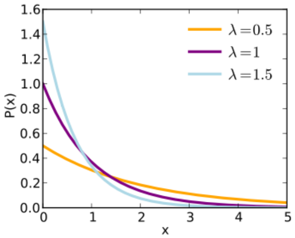
$$\begin{aligned} \lambda_{\mathrm{ML}} &= \arg\max_{\lambda} \log l(\lambda) \\ &= \arg\max_{\lambda} \left(\lambda e^{-\lambda 3.1} \lambda e^{-\lambda 8.2} \lambda e^{-\lambda 1.7}\right) \\ &= \arg\max_{\lambda} 3 \log \lambda + (-3.1 - 8.2 - 1.7)\lambda \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \lambda} \log l(\lambda) &= 3\frac{1}{\lambda} - 13 \\ & \rightarrow \lambda_{\mathrm{ML}} = \frac{3}{13} \end{aligned}$$



ML for Exponential Distribution

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$



- Consider having received samples $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

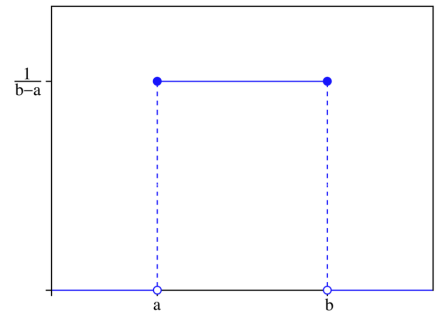
$$\begin{aligned} \log l(\lambda) &= \log \prod_{i=1}^{m} p(x^{(i)};\lambda) \\ &= \sum_{i=1}^{m} \log p(x^{(i)};\lambda) \\ &= \sum_{i=1}^{m} \log(\lambda e^{-\lambda x^{(i)}}) \\ &= \sum_{i=1}^{m} \log \lambda - \lambda x^{(i)} \\ &= m \log \lambda - \lambda \sum_{i=1}^{m} x^{(i)} \end{aligned}$$

$$\frac{\partial}{\partial \lambda} \log l(\lambda) = m\frac{1}{\lambda} - \sum_{i=1}^{m} x^{(i)}$$

$$\rightarrow \lambda_{\text{ML}} = \frac{1}{\frac{1}{m}\sum_{i=1}^{m} x^{(i)}}$$

Uniform

$$p(x; a, b) = \begin{cases} e^{-\lambda x}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases}$$



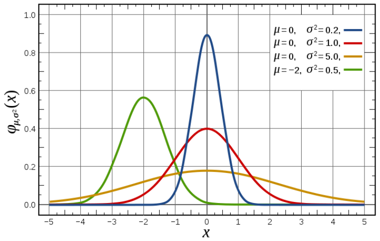
- Consider having received samples $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

$$\log l(a, b) = \sum_{i=1}^{m} \log \left(1\{x^{(i)} \in [a, b]\} \frac{1}{b - a} \right)$$

$$\rightarrow a_{\mathrm{ML}} = \min_i x^{(i)}, \quad b_{\mathrm{ML}} = \max_i x^{(i)}$$

ML for Gaussian

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- Consider having received samples $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

$$\begin{aligned}
\log l(\mu,\sigma) &= \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) \\
&= C + \sum_{i=1}^{m} -\log\sigma - \frac{(x^{(i)}-\mu)^2}{2\sigma^2}
\end{aligned}$$

$$\frac{\partial}{\partial\mu}\log l(\mu,\sigma) = \frac{1}{\sigma^2}\sum_{i=1}^{m}(x^{(i)}-\mu)$$

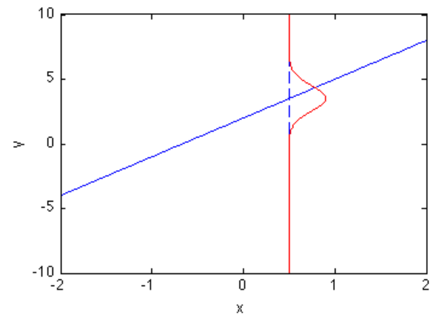
$$\rightarrow \mu_{\mathrm{ML}} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\frac{\partial}{\partial\sigma}\log l(\mu,\sigma) = \sum_{i=1}^{m} \frac{1}{\sigma} - \frac{(x^{(i)}-\mu)^2}{\sigma^3}$$

$$\rightarrow \sigma^2_{\mathrm{ML}} = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{\mathrm{ML}})^2$$

ML for Conditional Gaussian

$$y = a_0 + a_1 x + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$



Equivalently: $$p(y|x; a_0, a_1, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(a_0+a_1 x))^2}{2\sigma^2}}$$

More generally: $$y = a^\top x + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$p(y|x; a, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-a^\top x)^2}{2\sigma^2}}$$

ML for Conditional Gaussian

Given samples $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$.

$$\begin{aligned}
\log l(a, \sigma^2) &= \sum_{i=1}^{m} \log \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - a^\top x^{(i)})^2}{2\sigma^2}} \right) \\
&= C - m \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y^{(i)} - a^\top x^{(i)})^2
\end{aligned}$$

$$\begin{aligned}
\nabla_a \log l(a, \sigma^2) &= \frac{1}{\sigma^2} \sum_{i=1}^{m} (y^{(i)} - a^\top x^{(i)}) x^{(i)} \\
&= \sum_{i=1}^{m} y^{(i)} x^{(i)} - \left(\sum_{i=1}^{m} x^{(i)} x^{(i)\top} \right) a
\end{aligned}$$

$$\begin{aligned}
\rightarrow a_{\mathrm{ML}} &= \left(\sum_{i=1}^{m} x^{(i)} x^{(i)\top} \right)^{-1} \left(\sum_{i=1}^{m} y^{(i)} x^{(i)} \right) \\
&= (X^\top X)^{-1} X^\top y
\end{aligned}$$

$$X = \begin{bmatrix} x^{(1)\top} \\ x^{(2)\top} \\ \ldots \\ x^{(m)\top} \end{bmatrix} \qquad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \ldots \\ y^{(m)} \end{bmatrix}$$

$$\frac{\partial}{\partial \sigma} \log l(a, \sigma^2) = -m \frac{1}{\sigma} - \frac{1}{\sigma^3} \sum_{i=1}^{m} (y^{(i)} - a^\top x^{(i)})^2$$

$$\rightarrow \sigma^2_{\mathrm{ML}} = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - a^\top_{\mathrm{ML}} x^{(i)})^2$$

ML for Conditional Multivariate Gaussian

$$y = Cx + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \Sigma)$$

$$p(y|x; C, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{-1/2}} e^{-\frac{1}{2}(y-Cx)^\top \Sigma^{-1}(y-Cx)}$$

$$\log l(C, \Sigma) \quad = \quad -m\frac{n}{2}\log(2\pi) + \frac{m}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - Cx^{(i)})^\top \Sigma^{-1}(y^{(i)} - Cx^{(i)})$$

$$\nabla_{\Sigma^{-1}} \log l(C, \Sigma) \quad = \quad -\frac{m}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - C^\top x^{(i)})(y^{(i)} - C^\top x^{(i)})^\top$$

$$\rightarrow \qquad \Sigma_{\mathrm{ML}} = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - C^\top x^{(i)})(y^{(i)} - C^\top x^{(i)})^\top = \frac{1}{m}(Y^\top - CX^\top)(Y^\top - CX^\top)^\top$$

$$\nabla_C \log l(C, \Sigma) \quad = \quad -\frac{1}{2}\sum_{i=1}^{m}\Sigma^{-1}Cx^{(i)}x^{(i)\top} + x^{(i)}x^{(i)\top}C^\top\Sigma^{-1} - x^{(i)}y^{(i)\top}\Sigma^{-1} - \Sigma^{-1}y^{(i)}x^{(i)\top}$$

$$= \quad -\frac{1}{2}\left(\Sigma^{-1}CX^\top X + X^\top X C^\top \Sigma^{-1} - X^\top Y \Sigma^{-1} - \Sigma^{-1}Y^\top X\right)$$

$$\rightarrow \qquad C = Y^\top X (X^\top X)^{-1}$$

$$X = \begin{bmatrix} x^{(1)\top} \\ x^{(2)\top} \\ \dots \\ x^{(m)\top} \end{bmatrix} \qquad y = \begin{bmatrix} y^{(1)\top} \\ y^{(2)\top} \\ \dots \\ y^{(m)\top} \end{bmatrix}$$

Aside: Key Identities for Derivation on Previous Slide

$$\text{Trace}(A) = \sum_{i=1}^{n} A_{ii} \tag{1}$$

$$\text{Trace}(ABC) = \text{Trace}(BCA) = \text{Trace}(CAB) \tag{2}$$

$$\nabla_A \text{Trace}(AB) = B^\top \tag{3}$$

$$\nabla_A \log |A| = A^{-1} \tag{4}$$

Special case of (2), for $x \in \mathbb{R}^n$:

$$x^\top \Gamma x = \text{Trace}(x^\top \Gamma x) = \text{Trace}(\Gamma x x^\top) \tag{5}$$

ML Estimation in Fully Observed Linear Gaussian Bayes Filter Setting

- Consider the Linear Gaussian setting:

$$\begin{aligned} X_{t+1} &= AX_t + Bu_t + w_t \qquad w_t \sim \mathcal{N}(0,Q) \\ Z_{t+1} &= CX_t + d + v_t \qquad v_t \sim \mathcal{N}(0,R) \end{aligned}$$

- Fully observed, i.e., given $x_0, u_0, z_0, x_1, u_1, z_1, \ldots, x_T, u_T, z_t$

- → Two separate ML estimation problems for conditional multivariate Gaussian:

- 1:

$$X = \begin{bmatrix} x_0^\top u_0^\top \\ x_1^\top u_1^\top \\ \cdots \\ x_{T-1}^\top u_{T-1}^\top \end{bmatrix} \qquad y = \begin{bmatrix} x_1^\top \\ x_2^\top \\ \cdots \\ x_T^\top \end{bmatrix}$$

$$[A_{\mathrm{ML}} B_{\mathrm{ML}}] = Y^\top X (X^\top X)^{-1}$$

$$Q_{\mathrm{ML}} = \frac{1}{T} \sum_{t=0}^{T-1} (x_{t+1} - (Ax_t + Bu_t))(x_{t+1} - (Ax_t + Bu_t))^\top$$

- 2:

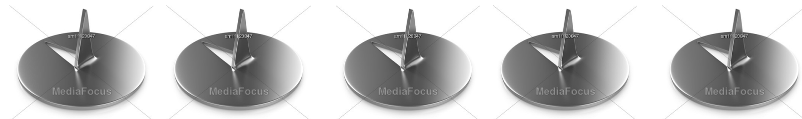
$$X = \begin{bmatrix} x_0^\top \\ x_1^\top \\ \cdots \\ x_T^\top \end{bmatrix} \qquad y = \begin{bmatrix} z_0^\top \\ z_1^\top \\ \cdots \\ z_T^\top \end{bmatrix}$$

$$[C_{\mathrm{ML}} d_{\mathrm{ML}}] = Y^\top X (X^\top X)^{-1}$$

$$R_{\mathrm{ML}} = \frac{1}{T} \sum_{t=0}^{T} (z_t - (Cx_t + d))(z_t - (Cx_t + d))^\top$$

Priors --- Thumbtack

- Let θ = P(up), 1-θ = P(down)



- How to determine θ ?

- ML estimate: 5 up, 0 down → $\theta_{\mathrm{ML}} = \frac{5}{5+0} = 1$

- Laplace estimate: add a fake count of 1 for each outcome

$$\theta_{\mathrm{Laplace}} = \frac{5+1}{5+1\ +\ 0+1} = \frac{6}{7}$$

Priors --- Thumbtack

- Alternatively, consider θ to be random variable
- Prior P(θ) = C θ(1-θ)
- Measurements: P(x | θ)

- Posterior:

$$\begin{aligned} P(\theta|x^{(1)},\ldots,x^{(5)}) &\propto P(\theta,x^{(1)},\ldots,x^{(5)}) \\ &= P(\theta)P(x^{(1)}|\theta)\ldots P(x^{(5)}|\theta) \\ &= \theta(1-\theta)\ \theta\theta\theta\theta\theta \\ &= \theta^6(1-\theta) \end{aligned}$$

- Maximum A Posterior (MAP) estimation
  - = find θ that maximizes the posterior

→ $\theta_{\text{MAP}} = \frac{6}{7}$

Priors --- Beta Distribution

$$P(\theta; \alpha, \beta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\theta_{\text{MAP}} = \frac{\alpha - 1 + n_1}{\alpha - 1 + n_1 + \beta - 1 + n_0}$$

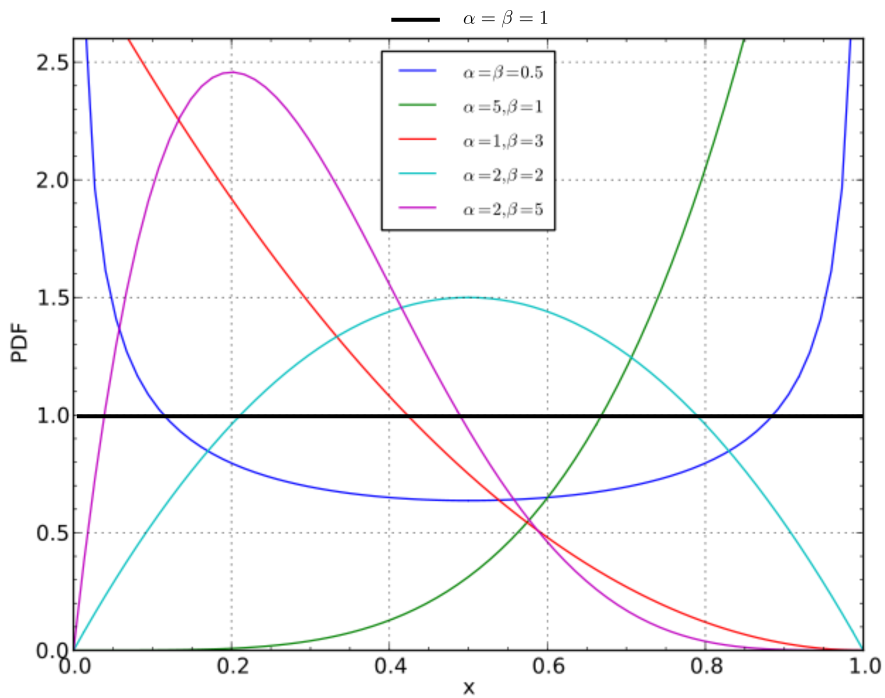


Figure source: Wikipedia

Priors --- Dirichlet Distribution

$$P(\theta; \alpha_1, \ldots, \alpha_K) = \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

$$\theta_k^{\mathrm{MAP}} = \frac{n_k + \alpha_k - 1}{\sum_{j=1}^{K}(n_j + \alpha_j - 1)}$$

- Generalizes Beta distribution
- MAP estimate corresponds to adding fake counts n_I, …, n_K

MAP for Mean of Univariate Gaussian

- Assume variance known. (Can be extended to also find MAP for variance.)
- **Prior:** $P(\mu; \mu_0, \sigma_0^2) = \mathcal{N}(\mu_0, \sigma_0^2)$

$$
\begin{aligned}
\log P(\mu; \mu_0, \sigma_0^2) + \log l(\mu) &= \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}\right) + \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}}\right) \\
&= C - \frac{(\mu-\mu_0)^2}{2\sigma_0^2} - \sum_{i=1}^{m} \frac{(x^{(i)}-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\frac{\partial}{\partial\mu}(\log P(\mu; \mu_0, \sigma_0) + \log l(\mu)) = \frac{1}{\sigma_0^2}(\mu_0 - \mu) + \frac{1}{\sigma^2}\sum_{i=1}^{m}(x^{(i)} - \mu)
$$

$$
\rightarrow \mu_{\mathrm{ML}} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{m} x^{(i)}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{m}{\sigma^2}}
$$

MAP for Univariate Conditional Linear Gaussian

- Assume variance known. (Can be extended to also find MAP for variance.)
- Prior: $P(a; \mu_0, \Sigma_0) = \mathcal{N}(\mu_0, \Sigma_0)$

$$\begin{aligned}
\log P(a; \mu_0, \Sigma_0) + \log l(a) &= \log\left(\frac{1}{(2\pi)^{n/2}|\Sigma_0|^{1/2}} e^{-\frac{1}{2}(a-\mu_0)^\top \Sigma_0^{-1}(a-\mu_0)}\right) + \sum_{i=1}^{m} \log\left(\frac{1}{(2\pi)^{1/2}\sigma} e^{-\frac{(a^\top x^{(i)} - y^{(i)})^2}{2\sigma^2}}\right) \\
&= C - \frac{1}{2}(a-\mu_0)^\top \Sigma_0^{-1}(a-\mu_0) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(a^\top x^{(i)} - y^{(i)})^2
\end{aligned}$$

$$\begin{aligned}
\nabla_a (\cdots) &= -\Sigma_0^{-1}(a-\mu_0) - \frac{1}{\sigma^2}\sum_{i=1}^{m}(a^\top x^{(i)} - y^{(i)})x^{(i)} \\
&= -(\Sigma_0^{-1} + \frac{1}{\sigma^2}X^\top X)a + \Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}X^\top y
\end{aligned}$$

$$\to a_{\mathrm{ML}} = (\Sigma_0^{-1} + \frac{1}{\sigma^2}X^\top X)^{-1}(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}X^\top y) \qquad X = \begin{bmatrix} x^{(1)\top} \\ x^{(2)\top} \\ \dots \\ x^{(m)\top} \end{bmatrix} \qquad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(m)} \end{bmatrix}$$

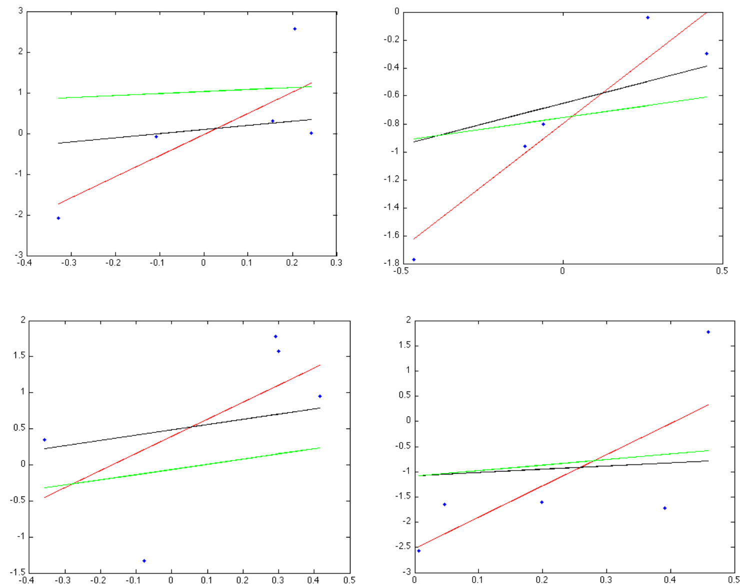
[Interpret!]

MAP for Univariate Conditional Linear Gaussian: Example

$$\mu_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sigma = 1$$

```
for run=1:4

    a = randn;
    b = randn;
    x = (rand(5,1) - 0.5);
    y = a*x + b + randn(5,1);
    X = [ones(5,1) x];
    ba_ML = (X'*X)^(-1)*X'*y;
    ba_MAP = (eye(2) + X'*X)^(-1)*(X'*y);
    figure; plot(x, y, '.');
    hold on;
    plot(x, ba_ML(1) + ba_ML(2)*x, 'r-');
    plot(x, ba_MAP(1) + ba_MAP(2)*x, 'k-');
    plot(x, b + a*x, 'g-');
end
```



TRUE ---

Samples .

ML ---

MAP ---

Cross Validation

- Choice of prior will heavily influence quality of result
- Fine-tune choice of prior through cross-validation:
  - 1. Split data into "training" set and "validation" set
  - 2. For a range of priors,
    - Train: compute θ_{MAP} on training set
    - Cross-validate: evaluate performance on validation set by evaluating the likelihood of the validation data under θ_{MAP} just found
  - 3. Choose prior with highest validation score
    - For this prior, compute θ_{MAP} on (training+validation) set
- Typical training / validation splits:
  - 1-fold: 70/30, random split
  - 10-fold: partition into 10 sets, average performance for each set being the validation set and the other 9 being the training set

Outline

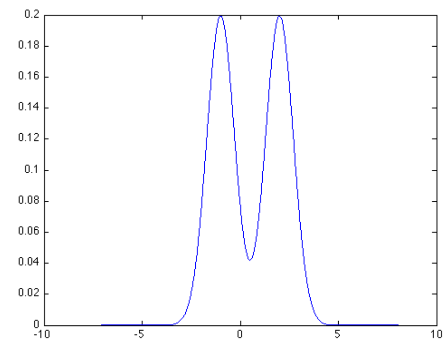
- Maximum likelihood (ML)
- Priors, and maximum a posteriori (MAP)
- Cross-validation
- **Expectation Maximization (EM)**

Mixture of Gaussians

- Generally:
$$\begin{aligned} X &\sim \text{Multinomial}(\theta) \\ Z|X=k &\sim \mathcal{N}(\mu_k, \Sigma_k) \end{aligned}$$

- Example:
$$P(X=1) = \frac{1}{2}, P(X=2) = \frac{1}{2}$$
$$Z|X=1 \sim \mathcal{N}(-1,1)$$
$$Z|X=2 \sim \mathcal{N}(2,1)$$

$$\rightarrow Z \sim \frac{1}{2}\mathcal{N}(-1,1) + \frac{1}{2}\mathcal{N}(2,1)$$



- ML Objective: given data $z^{(1)}, \ldots, z^{(m)}$

$$\max_{\theta,\mu,\Sigma} \sum_{i=1}^{m} \log \sum_{k=1}^{n} \theta_k \frac{1}{(2\pi)^{d/2}|\Sigma_k|} e^{-\frac{1}{2}(z-\mu_k)^\top \Sigma_k^{-1}(z-\mu_k)}$$

- Setting derivatives w.r.t. θ, μ, Σ equal to zero does not enable to solve for their ML estimates in closed form

We can evaluate function → we can in principle perform local optimization. In this lecture: "EM" algorithm, which is typically used to efficiently optimize the objective (locally)

Expectation Maximization (EM)

- Example:
  - Model: $$P(X=1)=\frac{1}{2}, P(X=2)=\frac{1}{2}$$ $$Z|X=1 \sim \mathcal{N}(\mu_1, 1)$$ $$Z|X=2 \sim \mathcal{N}(\mu_2, 1)$$
  - Goal:
    - Given data $z^{(1)}, \ldots, z^{(m)}$ (but no $x^{(i)}$ observed)
    - Find maximum likelihood estimates of μ_1, μ_2
  - EM basic idea: if $x^{(i)}$ were known → two easy-to-solve separate ML problems
  - EM iterates over
    - **E-step**: For i=1,…,m fill in missing data $x^{(i)}$ according to what is most likely given the current model [1]
    - **M-step**: run ML for completed data, which gives new model [1]

EM Derivation

- EM solves a Maximum Likelihood problem of the form:

$$\max_\theta \log \int_x p(x, z; \theta) dx$$

μ: parameters of the probabilistic model we try to find

x: unobserved variables

z: observed variables

$$\begin{aligned}
\max_\theta \log \int_x p(x, z; \theta) dx &= \max_\theta \log \int_x \frac{q(x)}{q(x)} p(x, z; \theta) dx \\
&= \max_\theta \log \int_x q(x) \frac{p(x, z; \theta)}{q(x)} dx \\
&= \max_\theta \log E_{X \sim q} \left[\frac{p(X, z; \theta)}{q(X)} \right] \\
\textit{Jensen's Inequality} \qquad &\geq \max_\theta E_{X \sim q} \log \left[\frac{p(X, z; \theta)}{q(X)} \right] \\
&= \max_\theta \int_x q(x) \log p(x, z; \theta) dx - \int_x q(x) \log q(x) dx
\end{aligned}$$

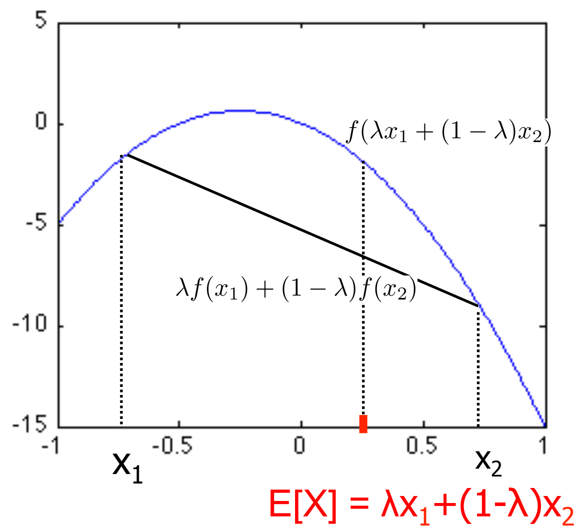
Jensen's inequality

Suppose f is concave, then for all probability measures P we have that:

$$f(\mathrm{E}_{X\sim P}) \geq E_{X\sim P}[f(X)]$$

with equality holding only if f is an affine function.

Illustration:
P(X=x_1) = 1-λ,
P(X=x_2) = λ



EM Derivation (ctd)

$$\max_\theta \log \int_x p(x,z;\theta)dx \geq \max_\theta \int_x q(x)\log p(x,z;\theta)dx - \int_x q(x)\log q(x)dx$$

Jensen's Inequality: equality holds when $f(x) = \log \frac{p(x,z;\theta)}{q(x)}$ *is an affine function.*

This is achieved for $q(x) = p(x|z;\theta) \propto p(x,z;\theta)$

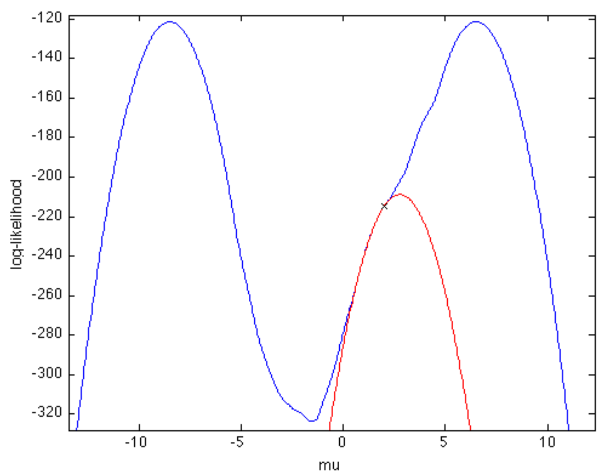
EM Algorithm: Iterate

1. E-step: Compute $q(x) = p(x|z;\theta)$
2. M-step: Compute $\theta = \arg\max_\theta \int_x q(x)\log p(x,z;\theta)dx$

M-step optimization can be done efficiently in most cases
E-step is usually the more expensive step
It does not fill in the missing data x with hard values, but finds a distribution q(x)

EM Derivation (ctd)

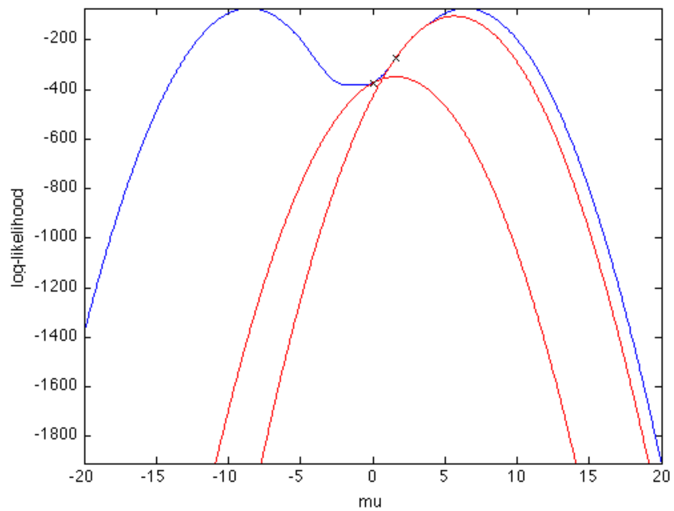
- M-step objective is upper-bounded by true objective
- M-step objective is equal to true objective at current parameter estimate



- → Improvement in true objective is at least as large as improvement in M-step objective

EM 1-D Example --- 2 iterations

- Estimate 1-d mixture of two Gaussians with unit variance:
  - $p(x;\mu) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu_1)^2} + \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu_2)^2}$



  - one parameter μ; $\mu_1 = \mu - 7.5$, $\mu_2 = \mu + 7.5$

EM for Mixture of Gaussians

- X ~ Multinomial Distribution, $P(X=k\ ;\ \theta) = \theta_k$
- Z ~ $N(\mu_k, \Sigma_k)$
- Observed: $z^{(1)}, z^{(2)}, \ldots, z^{(m)}$

$$p(x=k, z; \theta, \mu, \Sigma) = \theta_k \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(z-\mu_k)^\top \Sigma_k^{-1}(z-\mu_k)}$$

$$p(z; \theta, \mu, \Sigma) = \sum_{k=1}^{K} \theta_k \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(z-\mu_k)^\top \Sigma_k^{-1}(z-\mu_k)}$$

EM for Mixture of Gaussians

- E-step: $q(x) = p(x|z;\theta,\mu,\Sigma) = \prod_{i=1}^{m} p(x^{(i)}|z^{(i)};\theta,\mu,\Sigma)$

$$\begin{aligned} \rightarrow q(x^{(i)}=k) &= p(x^{(i)}=k|z^{(i)};\theta,\mu,\Sigma) \\ &\propto p(x^{(i)}=k, z^{(i)};\theta,\mu,\Sigma) \\ &= \theta_k \mathcal{N}(z^{(i)};\mu_k,\Sigma_k) \end{aligned}$$

- M-step: $\max_{\theta,\mu,\Sigma} \sum_{i=1}^{m}\sum_{k=1}^{k} q(x^{(i)}=k) \log\left(\theta_k \mathcal{N}(z^{(i)};\mu_k,\Sigma_k)\right)$

$$\rightarrow \theta_k = \frac{1}{m}\sum_{i=1}^{m} q(x^{(i)}=k) \qquad \rightarrow \mu_k = \frac{1}{\sum_{i=1}^{m} q(x^{(i)}=k)} q(x^{(i)}=k) z^{(i)}$$

$$\rightarrow \Sigma_k = \frac{1}{\sum_{i=1}^{m} q(x^{(i)}=k)} q(x^{(i)}=k)(z^{(i)}-\mu_k)(z^{(i)}-\mu_k)^\top$$

ML Objective HMM

- Given samples $\{z_0, z_1, z_2, \ldots, z_T\}, \ x_t \in \{1, 2, \ldots, I\}, z_t \in \{1, 2, \ldots, K\}$
- Dynamics model: $P(x_{t+1} = i | x_t = j) = \theta_{i|j}$
- Observation model: $P(z_t = k | z_t = l) = \gamma_{k|l}$
- ML objective:

$$
\begin{aligned}
\log l(\theta, \gamma) &= \log \left(\sum_{x_0, x_1, \ldots, x_T} P(x_0) \prod_{t=1}^{T} P(x_t | x_{t-1}; \theta) P(z_t | x_t; \gamma) \right) \\
&= \log \left(\sum_{x_0, x_1, \ldots, x_T} P(x_0) \prod_{t=1}^{T} \theta_{x_t | x_{t-1}} \prod_{t=1}^{T} \gamma_{z_t | x_t} \right)
\end{aligned}
$$

→ No simple decomposition into independent ML problems for each $\theta_{\cdot|j}$ and each $\gamma_{\cdot|l}$

→ No closed form solution found by setting derivatives equal to zero

EM for HMM --- M-step

$$
\begin{aligned}
& \max_{\theta,\gamma} \sum_{x_{0:T}} q(x_{0:T}) \log p(x_{0:T}, z_{0:T}; \theta, \gamma) \\
= \quad & \max_{\theta,\gamma} \sum_{x_{0:T}} q(x_{0:T}) \left(\sum_{t=0}^{T-1} \log p(x_{t+1}|x_t; \theta) + \sum_{t=0}^{T} \log p(z_t|x_t; \gamma) \right) \\
= \quad & \max_{\theta,\gamma} \sum_{t=0}^{T-1} \sum_{x_t, x_{t+1}} q(x_t, x_{t+1}) \log p(x_{t+1}|x_t; \theta) + \sum_{t=0}^{T} \sum_{x_t} q(x_t) \log p(z_t|x_t; \gamma)
\end{aligned}
$$

- θ and γ computed from "soft" counts

$$
\theta_{i|j} = \frac{n_{(i,j)}}{\sum_{i'=1}^{I} n_{(i',j)}} \qquad \gamma_{k|l} = \frac{m_{(k,l)}}{\sum_{k'=1}^{K} m_{(k',l)}}
$$

$$
\begin{aligned}
n_{(i,j)} &= \sum_{t=0}^{T-1} q(x_{t+1} = i, x_t = j) \\
m_{(k,l)} &= \sum_{t=0}^{T} q(z_t = k, x_t = l)
\end{aligned}
$$

EM for HMM --- E-step

- No need to find conditional full joint

$$q(x_{0:T}) = p(x_{0:T} | z_{0:T}; \theta, \gamma)$$

- Run smoother to find:

$$\begin{aligned} q(x_t, x_{t+1}) &= p(x_t, x_{t+1} | z_{0:T}; \theta, \gamma) \\ q(x_t) &= p(x_t | z_{0:T}; \theta, \gamma) \end{aligned}$$

ML Objective for Linear Gaussians

- Linear Gaussian setting:

$$\begin{aligned} X_{t+1} &= AX_t + Bu_t + w_t \qquad w_t \sim \mathcal{N}(0, Q) \\ Z_{t+1} &= CX_t + d + v_t \qquad v_t \sim \mathcal{N}(0, R) \end{aligned}$$

- Given $u_0, z_0, u_1, z_1, \ldots, u_T, z_t$

- ML objective:

$$\max_{Q,R,A,B,C,d} \log \int_{x_{0:T}} p(x_{0:T}, z_{0:T}; Q, R, A, B, C, d)$$

- EM-derivation: same as HMM

EM for Linear Gaussians --- E-Step

- Forward:

$$\begin{aligned}\mu_{t+1|0:t} &= A_t \mu_{t|0:t} + B_t u_t \\ \Sigma_{t+1|0:t} &= A_t \Sigma_{t|0:t} A_t^\top + Q_t\end{aligned}$$

$$\begin{aligned}K_{t+1} &= \Sigma_{t+1|0:t} C_{t+1}^\top (C_{t+1} \Sigma_{t+1|0:t} C_{t+1}^\top + R_{t+1})^{-1} \\ \mu_{t+1|0:t+1} &= \mu_{t+1|0:t} + K_{t+1}(z_{t+1} - (C_{t+1}\mu_{t+1|0:t} + d)) \\ \Sigma_{t+1|0:t+1} &= (I - K_{t+1} C_{t+1}) \Sigma_{t+1|0:t}\end{aligned}$$

- Backward:

$$\begin{aligned}\mu_{t|0:T} &= \mu_{t|0:t} + L_t(\mu_{t+1|0:T} - \mu_{t+1|0:t}) \\ \Sigma_{t|0:T} &= \Sigma_{t|0:t} + L_t(\Sigma_{t+1|0:T} - \Sigma_{t+1|0:t}) L_t^\top \\ L_t &= \Sigma_{t|0:t} A_t^\top \Sigma_{t+1|0:t}^{-1}\end{aligned}$$

EM for Linear Gaussians --- M-step

$$
\begin{aligned}
Q &= \frac{1}{T}\sum_{t=0}^{T-1}(\mu_{t+1|0:T} - A_t\mu_{t|0:T} - B_t u_t)(\mu_{t+1|0:T} - A_t\mu_{t|0:T} - B_t u_t)^\top \\
&\quad + A_t\Sigma_{t|0:T}A_t^\top + \Sigma_{t+1|0:T} - \Sigma_{t+1|0:T}L_t^\top A_t^\top - A_t L_t \Sigma_{t+1|0:T} \\
R &= \frac{1}{T+1}\sum_{t=0}^{T}(z_t - C_t\mu_{t|0:T} - d_t)(z_t - C_t\mu_{t|0:T} - d_t)^\top + C_t\Sigma_{t|0:T}C_t^\top
\end{aligned}
$$

[Updates for A, B, C, d. TODO: Fill in once found/derived.]

EM for Linear Gaussians --- The Log-likelihood

- When running EM, it can be good to keep track of the log-likelihood score --- it is supposed to increase every iteration

$$
\begin{aligned}
\log \prod_{t=1}^{T} p(z_{0:T}) &= \log \left(p(z_0) \prod_{t=1}^{T} p(z_t | z_{0:t-1}) \right) \\
&= \log p(z_0) + \sum_{t=1}^{T} \log p(z_t | z_{0:t-1})
\end{aligned}
$$

$$
Z_t | z_{0:t-1} \sim \mathcal{N}(\bar{\mu}_t, \bar{\Sigma}_t)
$$

$$
\begin{aligned}
\bar{\mu}_t &= C_t \mu_{t|0:t-1} + d_t \\
\bar{\Sigma}_t &= C_t \Sigma_{t|0:t-1} C_t^\top + R_t
\end{aligned}
$$

EM for Extended Kalman Filter Setting

- As the linearization is only an approximation, when performing the updates, we might end up with parameters that result in a lower (rather than higher) log-likelihood score
- → Solution: instead of updating the parameters to the newly estimated ones, interpolate between the previous parameters and the newly estimated ones. Perform a “line-search” to find the setting that achieves the highest log-likelihood score