Contents

1. Lagrange multipliers example
2. EM for exponential family
3. Application to EM for a mixture of Poisson
Lagrange multiplier example

Lagrange multipliers: optimise $E$ under constraint(s).

Example

Find Maximum Entropy distribution under constraints:

1. Constant mean $\lambda^{-1}$
2. Continuous, over positive values.
Need to optimize Entropy:

\[ H[p(\tau)] = - \int_0^\infty d\tau p(\tau) \log p(\tau) \]

Under constraints:

\[ \int_0^\infty d\tau p(\tau) = 1 \quad (1) \]

\[ \int_0^\infty d\tau p(\tau) \tau = \lambda^{-1} \quad (2) \]

Positive values: support = [0; \infty)
Define Lagrangian $\mathcal{L}$:

$$
\mathcal{L} = -\int d\tau p(\tau) \log p(\tau) + \gamma_1 \left( \int d\tau p(\tau) - 1 \right) + \gamma_2 \left( \int d\tau p(\tau)\tau - \lambda^{-1} \right)
$$

(3)

Get partial (functional) derivative:

$$
\frac{\partial \mathcal{L}}{\partial p(\tau)} = -\log p(\tau) - \frac{p(\tau)}{p(\tau)} + \gamma_1 + \gamma_2 \tau = 0
$$

$$
\iff p(\tau) = e^{\gamma_1 + \gamma_2 \tau^{-1}}
$$

(4)

Get partial derivatives for $\gamma_1$ and $\gamma_2$:

$$
\int d\tau p(\tau) = 1
$$
$$
\int d\tau p(\tau)\tau = \lambda^{-1}
$$
Put (4) back into the constraints.

Into (1): $\int d\tau p(\tau) = 1$

\[
\int d\tau e^{\gamma_1 + \gamma_2 \tau - 1} = 1
\]

$\iff e^{\gamma_1 - 1} \int d\tau e^{\gamma_2 \tau} = 1$

$\iff e^{\gamma_1 - 1} \frac{1}{\gamma_2} \left[ \int d\tau \gamma'_2 e^{-\gamma'_2 \tau} \right] = 1$  \hspace{1cm} \text{change of variable } -\gamma_2 = \gamma'_2

$\iff e^{\gamma_1 - 1} = \gamma'_2$  \hspace{1cm} \text{integral of exponential distr. } = 1 \hspace{1cm} (5)
Into (2):

\[ \int d\tau \tau e^{\gamma_1 + \gamma_2 \tau} = \lambda^{-1} \]

\[ \iff e^{\gamma_1 - 1} \int d\tau \tau e^{\gamma_2 \tau} = \lambda^{-1} \]

\[ \iff e^{\gamma_1 - 1} \frac{1}{\gamma_2'} \int d\tau \tau \gamma_2' e^{-\gamma_2' \tau} = \lambda^{-1} \]

\[ = \frac{1}{\gamma_2'} \]

\[ \iff e^{\gamma_1 - 1} \frac{1}{\gamma_2'} = \lambda^{-1} \]

change of variable \(-\gamma_2 = \gamma_2'\)

mean of exponential distr

Putting (5) in, change back to \(\gamma_2\)

\[ \iff \gamma_2 = -\lambda \]

(6)
Finally, put back everything into (4):

\[ p(\tau) = e^{\gamma_1 + \gamma_2 \tau - 1} = e^{\gamma_1 - 1} e^{\gamma_2 \tau} = \lambda e^{-\lambda \tau} \]

\( \Rightarrow \) Exponential distribution of rate \( \lambda \)
EM for Exponential family

Yet another form:

\[ p(x|\theta) = \exp \left( \theta^T s(x) \right) \frac{f(x)}{g(\theta)} \]

Partition function:

\[ g(\theta) = \int \exp \left( \theta^T s(x) \right) f(x) dx \]

Interesting because:

\[ \frac{\partial \log g(\theta)}{\partial \theta} = E_{p(x|\theta)}[s(x)] \]

also:

\[ \frac{\partial^2 \log g(\theta)}{\partial \theta^2} = Var_{p(x|\theta)}[s(x)] \]
Proof of (7).

Using:

\[
\frac{\partial g(\theta)}{\partial \theta} = \int s(x) \exp(\theta^T s(x)) f(x) dx
\]

We have:

\[
\frac{\partial \log g(\theta)}{\partial \theta} = \frac{\partial g(\theta)}{g(\theta)} = \int s(x) \frac{\exp(\theta^T s(x)) f(x)}{g(\theta)} dx
\]
\[
= E_{p(x|\theta)}[s(x)]
\]
Definition

EM steps for Exponential family:

\[ E: \quad q(y) = p(y|x, \theta) \]
\[ M: \quad E_{p(x,y|\theta)}[s(x,y)] = E_{q(y)}[s(x,y)] \]
Proof.

Let $F(q, \theta) = \log p(x|\theta) - KL[q(y)\|p(y|x, \theta)]$ and
$p(x, y|\theta) = \exp(\theta^T s(x, y)) \frac{f(x,y)}{g(\theta)}$

E: Optimise $F$ wrt $q$, $\theta$ fixed. KL=0 only when both sides equal.

M: Optimise $F$ wrt $\theta$, $q$ fixed.

$$\log p(x, y|\theta) = \theta^T s(x, y) + \log f(x, y) - \log g(\theta)$$

$$F(q, \theta) = E_{q(y)}[\log p(x, y|\theta)] + H[q]$$

$$F(q, \theta) = \theta^T E_{q(y)}[s(x, y)] + E_{q(y)}[\log f(x, y)] - E_{q(y)}[\log g(\theta)] + H[q]$$

$$\frac{\partial F}{\partial \theta} = E_{q(y)}[s(x, y)] - \frac{\partial \log g(\theta)}{\partial \theta} = 0$$

Using (7): $$\frac{\partial \log g(\theta)}{\partial \theta} = E_{p(x|\theta)}[s(x)]$$

$$E_{p(x,y|\theta)}[s(x, y)] = E_{q(y)}[s(x, y)]$$
EM for Mixture of Poisson

Model:

\[ y \sim \text{Discrete}(\pi) \]
\[ x|y \sim \text{Poisson}(\lambda_y) \]

With

\[ x \in \mathbb{N} \quad y \in 1 \ldots M \]
\[ \pi = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_M \end{bmatrix} \quad \lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_M \end{bmatrix} \]

\[ p(y = k) = \pi_k \]

And

\[ p(x|y) = \frac{e^{-\lambda_y} \lambda_y^x}{x!} \]
\[ E_{p(x|y)}[x] = \lambda_y \]
E-step:

\[ q(y) = p(y|x, \theta) \]

\[ q(y) = p(y|x, \theta) \propto p(x, y|\theta) \]
\[ = p(y)p(x|y) \]
\[ = \pi_y \frac{e^{-\lambda_y \lambda_x^y}}{x!} \]
\[ \propto \pi_y e^{-\lambda_y \lambda_x^y} \]

\[ q(y) = P(y|x, \theta) = \frac{\pi_y e^{-\lambda_y \lambda_x^y}}{\sum_{y'=1}^{M} \pi_{y'} e^{-\lambda_{y'} \lambda_x^{y'}}} \]
M-step:

\[ E_{p(x,y|\theta)}[s(x, y)] = E_{q(y)}[s(x, y)] \]

Need \( s(x, y) \) and \( \theta \), a bit more work needed.
Idea:

\[ \log p(x, y|\theta) = \log \pi_y - \lambda_y + x \log \lambda_y - \log x! \]

Hard to get \( y \) out...
Instead:

\[ p(x, y|\theta) = \pi_y \frac{e^{-\lambda_y} \lambda_y^x}{x!} \]

\[ = \prod_{k=1}^{M} \left( \pi_k \frac{e^{-\lambda_k} \lambda_k^x}{x!} \right)^{\delta(y-k)} \]

Kronecker delta, selects only one term.
So now:

\[
\log p(x, y|\theta) = \sum_{k=1}^{M} \delta(y - k) [\log \pi_k - \lambda_k + x \log \lambda_k - \log x!]
\]

\[
= \sum_{k=1}^{M} \delta(y - k) [\log \pi_k - \lambda_k] + \sum_{k=1}^{M} x\delta(y - k) \log \lambda_k - \log x! \quad (8)
\]

Identify \(s(x, y)\) and \(\theta\):

\[
\begin{bmatrix}
\delta(y - 1) \\
\vdots \\
\delta(y - M) \\
x\delta(y - 1) \\
\vdots \\
x\delta(y - M)
\end{bmatrix} \quad \begin{bmatrix}
\log \pi_1 - \lambda_1 \\
\vdots \\
\log \pi_M - \lambda_M \\
\log \lambda_1 \\
\vdots \\
\log \lambda_M
\end{bmatrix}
\]

\((\log x! \text{ goes into } f(x, y))\)
All set to use the M-step identity!

\[ E_{p(x,y|\theta)}[s(x,y)] = E_{q(y)}[s(x,y)] \]

\[ E_{p(x,y|\theta)}[\delta(y - k)] = \sum_{y=1}^{M} \sum_{x=0}^{\infty} \delta(y - k) \underbrace{p(y|\theta)p(x|y,\theta)}_{=1} \]

\[ = \sum_{y=1}^{M} \delta(y - k) \pi_y \sum_{x=0}^{\infty} p(x|y) = \pi_k \]

\[ E_{q(y)}[\delta(y - k)] = \sum_{y=1}^{M} q(y) \delta(y - k) = q(k) \]

Using the identify:

\[ \pi_k = q(k) \]
And now for the rest of the sufficient statistics:

\[ E_{p(x,y|\theta)}[x\delta(y-k)] = \sum_{y=1}^{M} \sum_{x=0}^{\infty} x\delta(y-k)\pi_y p(x|y) \]

\[ = \sum_{y=1}^{M} \delta(y-k)\pi_y \left( \sum_{x=0}^{\infty} x p(x|y) \right) \]

\[ = \lambda_k \pi_k \]

\[ E_{q(y)}[x\delta(y-k)] = x E_{q(y)}[\delta(y-k)] = xq(k) \]

Using the identity:

\[ \pi_k \lambda_k = x q(k) \Rightarrow \lambda_k = \frac{xq(k)}{\pi_k} = x \]
EM for a Mixture of Poisson

E-step:

\[ q(y) = P(y|x, \theta) = \frac{\pi_y e^{-\lambda_y} \lambda_y^x}{\sum_{y'=1}^{M} \pi_{y'} e^{-\lambda_{y'}} \lambda_{y'}^x} \]

M-step:

\[
\begin{align*}
\pi_k &= q(k) \\
\lambda_k &= x
\end{align*}
\]

But this is only for 1 datapoint \( x \) . . .
Extension to $n$ iid datapoints.

$$p(\{x_i\}|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \exp(\theta^T \sum_{i=1}^{n} s(x_i)) \frac{\prod_{i=1}^{n} f(x_i)}{g(\theta)^n}$$

Exponential family is easy!
New sufficient statistics to match:

$$s(x, y) = \sum_{i=1}^{n} s(x_i, y_i) = 
\begin{bmatrix}
\sum_{i=1}^{n} \delta(y_i - 1) \\
\vdots \\
\sum_{i=1}^{n} \delta(y_i - M) \\
\sum_{i=1}^{n} x_i \delta(y_i - 1) \\
\vdots \\
\sum_{i=1}^{n} x_i \delta(y_i - M)
\end{bmatrix}$$

Recompute E and M updates!
E-step:

\[ q(y) = p(y|x, \theta) \]

\[ q(y) \propto p(x, y|\theta) \]
\[ = p(y)p(x|y) \]
\[ = \prod_{i=1}^{n} \pi_{y_i} \prod_{i=1}^{n} \frac{e^{-\lambda y_i} \lambda^{x_i}}{x_i!} \]
\[ \propto \prod_{i=1}^{n} \pi_{y_i} e^{-\lambda y_i} \lambda^{x_i} \]

As \( y_i \) are independent of each others.
Now this also means that \( q(y) \) factorises, as can be seen if we were to compute the marginals \( q_i(y_i) \):
\[ q_i(y_i) = \int dy_1 \ldots dy_{i-1} dy_{i+1} \ldots dy_n \prod_{k=1}^{n} \pi_{y_k} e^{-\lambda y_k} \lambda^{x_k}_{y_k} \]

\[ = \pi_{y_i} e^{-\lambda y_i} \lambda^{x_i}_{y_i} \int dy_1 \ldots dy_{i-1} dy_{i+1} \ldots dy_n \prod_{k \neq i} \pi_{y_k} e^{-\lambda y_k} \lambda^{x_k}_{y_k} \]

constant, doesn’t depend on \( y_i \)

\[ \propto \pi_{y_i} e^{-\lambda y_i} \lambda^{x_i}_{y_i} \]

\[ \Rightarrow q(y) = \prod_{i=1}^{n} q_i(y_i) \]

So we have our E-step by normalising each of them:

\[ q_i(y_i) = \frac{\pi_{y_i} e^{-\lambda y_i} \lambda^{x_i}_{y_i}}{\sum_y \pi_y e^{-\lambda y} \lambda^x_y} \]
Note: having each $q_i(y_i)$ normalised also ensures that $q(y)$ is normalised:

$$\sum_y q(y) = \sum_{y_1} \cdots \sum_{y_n} q(y)$$

$$= \sum_{y_1} \cdots \sum_{y_n} \prod_{i=1}^n q_i(y_i)$$

$$= \sum_{y_1} q_1(y_1) \left( \sum_{y_2} q_2(y_2) \left( \cdots \sum_{y_{n-1}} q_{n-1}(y_{n-1}) \left( \sum_{y_n} q_n(y_n) \right) \right) \right) \right)$$

$$= 1$$

(recursion from the right, as all sums are equal to 1)
M-step for general $n$:

$$E_{p(x, y|\theta)} \left[ \sum_{i=1}^{n} \delta(y_i - k) \right] = \sum_{x_1 \ldots x_n} \sum_{y_1 \ldots y_n} \left( \sum_{i=1}^{n} \delta(y_i - k) \right) \prod_{i=1}^{n} \pi_{y_i} p(x_i|y_i, \theta)$$

$$= \sum_{y_1 \ldots y_n} \left( \sum_{i=1}^{n} \delta(y_i - k) \right) \prod_{i=1}^{n} \pi_{y_i} \sum_{x_1 \ldots x_n} \prod_{i=1}^{n} p(x_i|y_i, \theta)$$

$$= \sum_{i=1}^{n} \sum_{y_i} \delta(y_i - k) \pi_{y_i}$$

$$= \sum_{i=1}^{n} \pi_k = n\pi_k \tag{9}$$

Where we use the fact that $p(y_i)$ factorises, similar to the previous slide.
\[
E_{q(y)} \left[ \sum_{i=1}^{n} \delta(y_i - k) \right] = \sum_{y_1 \ldots y_n} \left( \sum_{i=1}^{n} \delta(y_i - k) \right) \prod_{i=1}^{n} q_i(y_i)
\]

\[
= \sum_{i=1}^{n} \sum_{y_i} \delta(y_i - k) q_i(y_i)
\]

\[
= \sum_{i=1}^{n} q_i(k)
\]

(10)

Now putting (9) and (10) together:

\[
n\pi_k = \sum_{i=1}^{n} q_i(k) \Rightarrow \pi_k = \frac{1}{n} \sum_{i=1}^{n} q_i(k)
\]
Now for the rest of the sufficient statistics:

\[
\begin{align*}
E_{p(x,y|\theta)} \left[ \sum_{i=1}^{n} x_i \delta(y_i - k) \right] &= \sum_{x_1 \ldots x_n, y_1 \ldots y_n} \left( \sum_{i=1}^{n} x_i \delta(y_i - k) \right) \prod_{i=1}^{n} \pi_{y_i} p(x_i|y_i, \theta) \\
&= \sum_{i=1}^{n} \sum_{y_i} \delta(y_i - k) \pi_{y_i} \sum_{x_i} x_i p(x_i|y_i, \theta) \\
&= \sum_{i=1}^{n} \sum_{y_i} \delta(y_i - k) \pi_{y_i} \lambda_{y_i} = \sum_{i=1}^{n} \pi_k \lambda_k \\
&= n \lambda_k \pi_k = n \lambda_k \frac{1}{n} \sum_{i=1}^{n} q_i(k) = \lambda_k \sum_{i=1}^{n} q_i(k) \tag{11}
\end{align*}
\]

Where we used (10) on the last line.
$$E_{q(y)} \left[ \sum_{i=1}^{n} x_i \delta(y_i - k) \right] = \sum_{y_1 \ldots y_n} \left( \sum_{i=1}^{n} x_i \delta(y_i - k) \right) \prod_{i=1}^{n} q_i(y_i)$$

$$= \sum_{i=1}^{n} x_i \sum_{y_i} \delta(y_i - k) q_i(y_i)$$

$$= \sum_{i=1}^{n} x_i q_i(k) \quad (12)$$

Finally putting (11) and (12) together:

$$\lambda_k \sum_{i=1}^{n} q_i(k) = \sum_{i=1}^{n} x_i q_i(k) \Rightarrow \lambda_k = \frac{\sum_{i=1}^{n} x_i q_i(k)}{\sum_{i=1}^{n} q_i(k)}$$

Which ends the derivation of the M-step!
EM for Mixture of Poisson (solution)

E-step:

\[ q_i(y_i) = P(y_i|x_i, \theta) = \frac{\pi_y e^{-\lambda y_i} \lambda x y_i}{\sum_{y'=1}^{M} \pi_{y'} e^{-\lambda y'} \lambda x y'} \]

M-step:

\[
\begin{align*}
\pi_k &= \frac{\sum_{i=1}^{n} q_i(k)}{n} \\
\lambda_k &= \frac{\sum_{i=1}^{n} x_i q_i(k)}{\sum_{i=1}^{n} q_i(k)}
\end{align*}
\]