Tutorial 3
Lagrange multipliers and EM for exponential family

Loic Matthey

Gatsby Computational Neuroscience Unit

October 20th, 2010

Contents

Lagrange multiplier example

Lagrange multipliers: optimise E under constraint(s).

Example

Find Maximum Entropy distribution under constraints:

1. Constant mean λ^{-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:

$$\int_0^\infty d\tau p(\tau) = 1 \qquad (1)$$

$$\int_0^\infty d\tau p(\tau)\tau = \lambda^{-1} \qquad (2)$$

Positive values: support $= [0; \infty)$

Define Lagrangian $\mathcal{L}$:

$$\mathcal{L} = -\int d\tau p(\tau) \log p(\tau) + \gamma_1 \left(\int d\tau p(\tau) - 1 \right) + \gamma_2 \left(\int d\tau p(\tau)\tau - \lambda^{-1} \right) \quad (3)$$

Get partial (functional) derivative:

$$\frac{\partial L}{\partial p(\tau)} = -logp(\tau) - \underbrace{\frac{p(\tau)}{p(\tau)}}_{1} + \gamma_1 + \ \gamma_2\tau = 0$$

$$\iff \ p(\tau) = e^{\gamma_1 + \gamma_2 \tau - 1} \quad (4)$$

Get partial derivatives for γ_1 and γ_2:

$$\int d\tau p(\tau) = 1 \qquad\qquad \int d\tau p(\tau)\tau = \lambda^{-1}$$

Put (4) back into the constraints.

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

$$
\begin{aligned}
& \int d\tau e^{\gamma_1 + \gamma_2 \tau - 1} = 1 \\
\iff & e^{\gamma_1 - 1} \int d\tau e^{\gamma_2 \tau} = 1 \\
\iff & e^{\gamma_1 - 1} \frac{1}{\gamma_2'} \underbrace{\int d\tau \ \gamma_2' e^{-\gamma_2' \tau}}_{=1} = 1 && \text{change of variable } -\gamma_2 = \gamma_2' \\
\iff & e^{\gamma_1 - 1} = \gamma_2' && \text{integral of exponential distr. } = 1 \qquad (5)
\end{aligned}
$$

Into (2):

$$
\begin{aligned}
&\int d\tau \tau e^{\gamma_1+\gamma_2\tau-1} = \lambda^{-1} \\
\Longleftrightarrow\ & e^{\gamma_1-1}\int d\tau \tau e^{\gamma_2\tau} = \lambda^{-1} \\
\Longleftrightarrow\ & e^{\gamma_1-1}\frac{1}{\gamma_2'}\underbrace{\int d\tau\ \tau\ \gamma_2' e^{-\gamma_2'\tau}}_{=\frac{1}{\gamma_2'}} = \lambda^{-1} && \text{change of variable } -\gamma_2 = \gamma_2' \\
\Longleftrightarrow\ & e^{\gamma_1-1}\frac{1}{\gamma_2'^2} = \lambda^{-1} && \text{mean of exponential distr} \\
\Longleftrightarrow\ & \gamma_2'\frac{1}{\gamma_2'^2} = \lambda^{-1} && \text{Putting (5) in, change back to } \gamma_2 \\
\Longleftrightarrow\ & \gamma_2 = -\lambda && (6)
\end{aligned}
$$

Finally, put back everything into (4):

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

$$= \underbrace{e^{\gamma_1 - 1}}_{\lambda} e^{\gamma_2 \tau}$$

$$= \lambda e^{-\lambda \tau}$$

$\Rightarrow$ Exponential distribution of rate λ

EM for Exponential family

Yet another form:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T \mathbf{s}(\mathbf{x})\right) \frac{f(\mathbf{x})}{g(\boldsymbol{\theta})}$$

Partition function:

$$g(\theta) = \int \exp\left(\boldsymbol{\theta}^T \mathbf{s}(\mathbf{x})\right) f(\mathbf{x}) d\mathbf{x}$$

Interesting because:

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

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:

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

$$= E_{p(x|\theta)}[s(x)]$$

$\square$

Definition

EM steps for Exponential family:

$$\begin{aligned} \text{E:} \quad & q(y) = p(y|x, \theta) \\ \text{M:} \quad & E_{p(x,y|\theta)}[s(x,y)] = E_{q(y)}[s(x,y)] \end{aligned}$$

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, θ fixed. KL=0 only when both sides equal.

M: Optimise F wrt θ, 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:

$$y \sim \textit{Discrete}(\boldsymbol{\pi})$$
$$x|y \sim \textit{Poisson}(\lambda_y)$$

With

$$x \in \mathbb{N} \qquad y \in 1 \ldots M \qquad \boldsymbol{\pi} = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_M \end{bmatrix} \qquad \boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_M \end{bmatrix}$$

$$p(y = k) = \pi_k$$

And

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

E-step:

$$q(y) = p(y|x,\theta)$$

$$\begin{aligned} q(y) = p(y|x,\theta) &\propto p(x,y|\theta) \\ &= p(y)p(x|y) \\ &= \pi_y \frac{e^{-\lambda_y}\lambda_y^x}{x!} \\ &\propto \pi_y e^{-\lambda_y}\lambda_y^x \end{aligned}$$

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

M-step:

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

Need $s(x,y)$ and θ, 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:

$$\begin{aligned}\log p(x, y|\theta) &= \sum_{k=1}^{M} \delta(y-k)\left[\log \pi_k - \lambda_k + x \log \lambda_k - \log x!\right] \\ &= \sum_{k=1}^{M} \delta(y-k)\left[\log \pi_k - \lambda_k\right] + \sum_{k=1}^{M} x\delta(y-k) \log \lambda_k - \log x! \qquad (8)\end{aligned}$$

Identify $s(x, y)$ and θ:

$$s(x,y) = \begin{bmatrix} \delta(y-1) \\ \vdots \\ \delta(y-M) \\ x\delta(y-1) \\ \vdots \\ x\delta(y-M) \end{bmatrix} \qquad \theta = \begin{bmatrix} \log \pi_1 - \lambda_1 \\ \vdots \\ \log \pi_M - \lambda_M \\ \log \lambda_1 \\ \vdots \\ \log \lambda_M \end{bmatrix}$$

($\log x!$ goes into $f(x, y)$)

All set to use the M-step identity!

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

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

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

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

Using the identify:

$$\pi_k = q(k)$$

And now for the rest of the sufficient statistics:

$$
\begin{aligned}
E_{p(x,y|\theta)}[x\delta(y-k)] &= \sum_{y=1}^{M}\sum_{x=0}^{\infty} x\delta(y-k)\pi_y p(x|y) \\
&= \sum_{y=1}^{M}\delta(y-k)\pi_y \underbrace{\sum_{x=0}^{\infty} x\ p(x|y)}_{E_{p(x|y)}[x]=\lambda_y} \\
&= \lambda_k \pi_k
\end{aligned}
$$

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

Using the identify:

$$
\pi_k \lambda_k = x\ q(k) \Rightarrow \lambda_k = \frac{xq(k)}{\pi_k} = x
$$

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:

$$\begin{cases} \pi_k = q(k) \\ \lambda_k = x \end{cases}$$

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:

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

Recompute E and M updates!

E-step:

$$q(\mathbf{y}) = p(\mathbf{y}|\mathbf{x}, \theta)$$

$$\begin{aligned} q(\mathbf{y}) &\propto p(\mathbf{x}, \mathbf{y}|\theta) \\ &= p(\mathbf{y})p(\mathbf{x}|\mathbf{y}) \\ &= \prod_{i=1}^{n} \pi_{y_i} \prod_{i=1}^{n} \frac{e^{-\lambda_{y_i}} \lambda_{y_i}^{x_i}}{x_i!} \\ &\propto \prod_{i=1}^{n} \pi_{y_i} e^{-\lambda_{y_i}} \lambda_{y_i}^{x_i} \end{aligned}$$

As y_i are independent of each others.
Now this also means that $q(\mathbf{y})$ factorises, as can be seen if we were to compute the marginals $q_i(y_i)$:

$$
\begin{aligned}
q_i(y_i) &= \int dy_1 \dots dy_{i-1} dy_{i+1} \dots dy_n \prod_{k=1}^{n} \pi_{y_k} e^{-\lambda_{y_k}} \lambda_{y_k}^{x_k} \\
&= \pi_{y_i} e^{-\lambda_{y_i}} \lambda_{y_i}^{x_i} \underbrace{\int dy_1 \dots dy_{i-1} dy_{i+1} \dots dy_n \prod_{k \neq i} \pi_{y_k} e^{-\lambda_{y_k}} \lambda_{y_k}^{x_k}}_{\text{constant, doesn't depend on } y_i} \\
&\propto \pi_{y_i} e^{-\lambda_{y_i}} \lambda_{y_i}^{x_i} \\
&\Rightarrow q(\mathbf{y}) = \prod_{i=1}^{n} q_i(y_i)
\end{aligned}
$$

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

$$
q_i(y_i) = \frac{\pi_{y_i} e^{-\lambda_{y_i}} \lambda_{y_i}^{x_i}}{\sum_y \pi_y e^{-\lambda_y} \lambda_y^x}
$$

Note: having each $q_i(y_i)$ normalised also ensures that $q(\mathbf{y})$ is normalised:

$$
\begin{aligned}
\sum_{\mathbf{y}} q(\mathbf{y}) &= \sum_{y_1} \cdots \sum_{y_n} q(\mathbf{y}) \\
&= \sum_{y_1} \cdots \sum_{y_n} \prod_{i=1}^{n} q_i(y_i) \\
&= \sum_{y_1} q_1(y_1) \left(\sum_{y_2} q_2(y_2) \left(\cdots \sum_{y_{n-1}} q_{n-1}(y_{n-1}) \underbrace{\left(\sum_{y_n} q_n(y_n) \right)}_{=1\text{, as normalised}} \right) \right) \\
&= 1
\end{aligned}
$$

(recursion from the right, as all sums are equal to 1)

M-step for general n:

$$
\begin{aligned}
E_{p(\mathbf{x},\mathbf{y}|\theta)}\left[\sum_{i=1}^{n}\delta(y_i-k)\right] &= \sum_{x_1\ldots x_n}\sum_{y_1\ldots y_n}\left(\sum_{i=1}^{n}\delta(y_i-k)\right)\prod_{i=1}^{n}\pi_{y_i}p(x_i|y_i,\theta)\\
&= \sum_{y_1\ldots y_n}\left(\sum_{i=1}^{n}\delta(y_i-k)\right)\prod_{i=1}^{n}\pi_{y_i}\underbrace{\sum_{x_1\ldots x_n}\prod_{i=1}^{n}p(x_i|y_i,\theta)}_{=1}\\
&= \sum_{i=1}^{n}\sum_{y_i}\delta(y_i-k)\pi_{y_i}\\
&= \sum_{i=1}^{n}\pi_k = n\pi_k
\end{aligned}
\tag{9}
$$

Where we use the fact that $p(y_i)$ factorises, similar to the previous slide.

$$
\begin{aligned}
E_{q(\mathbf{y})}\left[\sum_{i=1}^{n}\delta(y_i - k)\right] &= \sum_{y_1\ldots y_n}\left(\sum_{i=1}^{n}\delta(y_i - k)\right)\prod_{i=1}^{n} q_i(y_i) \\
&= \sum_{i=1}^{n}\sum_{y_i}\delta(y_i - k)q_i(y_i) \\
&= \sum_{i=1}^{n} q_i(k) \qquad (10)
\end{aligned}
$$

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:

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

$$= \sum_{i=1}^{n}\sum_{y_i}\delta(y_i - k)\pi_{y_i}\underbrace{\sum_{x_i} x_i p(x_i|y_i,\theta)}_{=\lambda_{y_i}}$$

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

$$= n\lambda_k\pi_k = n\lambda_k\frac{1}{n}\sum_{i=1}^{n} q_i(k) = \lambda_k\sum_{i=1}^{n} q_i(k) \qquad (11)$$

Where we used (10) on the last line.

$$
\begin{aligned}
E_{q(\mathbf{y})}\left[\sum_{i=1}^{n} x_i \delta(y_i - k)\right] &= \sum_{y_1 \dots y_n} \left(\sum_{i=1}^{n} x_i \delta(y_i - k)\right) \prod_{i=1}^{n} q_i(y_i) \\
&= \sum_{i=1}^{n} x_i \sum_{y_i} \delta(y_i - k) q_i(y_i) \\
&= \sum_{i=1}^{n} x_i q_i(k) \qquad (12)
\end{aligned}
$$

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:

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

M-step:

$$\begin{cases} \pi_k &= \frac{\sum_{i=1}^{n} q_i(k)}{n} \\ \lambda_k &= \frac{\sum_{i=1}^{n} x_i q_i(k)}{\sum_{i=1}^{n} q_i(k)} \end{cases}$$