Problem Sheet 2

```
begin
    using Pkg; Pkg.add(["Distributions", "LinearAlgebra", "Plots", "PlutoUI"])
    using Distributions
    using LinearAlgebra
    using Plots
    using PlutoUI
    default(;linewidth=3.0, legendfontsize=15.0)
end
```

Table of Contents

1. EM algorithm for a Poisson mixture model

Consider a mixture model for a integer valued random variable $n \in \{0, 1, 2, \ldots\}$ given by the distribution

$$P(n|\boldsymbol{\theta}) = \sum_{j=1}^{M} P(j)\ P(n|\theta_j) = \sum_{j=1}^{M} P(j)\ e^{-\theta_j} \frac{\theta_j^n}{n!}\ ,$$

where the component probabilities $P(n|\theta_j)$ are Poisson distributions. Based on a data set of i.i.d.~samples $D = (n_1, n_2, \ldots, n_N)$ we want to estimate the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M, P(1), \ldots, P(M))$ of this mixture model.

(a) [MATH] Derive an expression for the Maximum Likelihood estimate of θ_1 for $M = 1$, where all obervations come from the same Poisson distribution.

Solution

- Likelihood of the data set:

$$P(D|\theta_1) = \prod_{i=1}^{N} P(n_i|\theta_1) = \prod_{i=1}^{N} \exp(-\theta_1) \frac{\theta_1^{n_i}}{n_i!} = \exp(-N\theta_1) \prod_{i=1}^{N} \frac{\theta_1^{n_i}}{n_i!}$$

- Logarithm of the likelihood:

$$F = -\log P(D|\theta_1) = N\theta_1 - \sum_{i=1}^{N} n_i \log \theta_1 + \sum_{i=1}^{N} \log n_i!$$

- Calculation of the Maximum-Likelihood estimate:

$$\left.\frac{dF}{d\theta_1}\right|_{\theta_1=\theta^*} = 0 \quad \Longleftrightarrow \quad N - \sum_{i=1}^{N} \frac{n_i}{\theta^*} = 0 \quad \Longleftrightarrow \quad \theta^* = \frac{1}{N} \sum_{i=1}^{N} n_i$$

(b) [MATH] For $M > 1$ the maximum likelihood estimates of the parameters are to be determined using an EM algorithm. Give explicit formulas for the update of θ_j and $P(j)$.

Tip

For the E-step (see the lecture), compute

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = -\sum_{i=1}^{N} \sum_{j=1}^{M} P_t(j|n_i, \theta_t) \ln \left(P(n_i|\theta_j)\, P(j) \right),$$

where $P_t(j|n_i)$ is the responsibility of component j for generating data point n_i, computed with the current values of the parameters. For the M-step, minimise $\mathcal{L}$ with respect to θ_j and $P(j)$.

Solution

- We need to compute the expectation given the posterior of $P_t(j|n_i, \theta_t)$:the posterior for generating observation n_i from component j of the mixture model is

$$P_t(j|n_i, \theta_t) = \left. \frac{P(j) e^{-\theta_j} \frac{\theta_j^{n_i}}{n_i!}}{\sum_{k=1}^{M} P(k) e^{-\theta_k} \frac{\theta_k^{n_i}}{n_i!}} \right|_{\theta=\theta_t} = \left. \frac{P(j) e^{-\theta_j} \theta_j^{n_i}}{\sum_{k=1}^{M} P(k) e^{-\theta_k} \theta_k^{n_i}} \right|_{\theta=\theta_t}$$

- E step: The expected log-likelihood is then given by :

$$\begin{aligned} \langle \mathcal{L} \rangle &= -\sum_{i=1}^{N} \sum_{j=1}^{M} P_t(j|n_i) \ln \left(P(j) e^{-\theta_j} \frac{\theta_j^{n_i}}{n_i!} \right) \\ &= -\sum_{i=1}^{N} \sum_{j=1}^{M} P_t(j|n_i) (-\theta_j + n_i \ln \theta_j - \ln n_i! + \ln P(j)) \end{aligned}$$

- M step: We can now maximixe $\langle \mathcal{L} \rangle$ given our parameters:

$$\begin{aligned} \frac{\partial \langle \mathcal{L} \rangle}{\partial \theta_j} = 0 &\iff -\sum_{i=1}^{N} P_t(j|n_i) \left(-1 + \frac{n_i}{\theta_j} \right) = 0 \\ &\iff \theta_j = \frac{\sum_{i=1}^{N} n_i P_t(j|n_i)}{\sum_{i=1}^{N} P_t(j|n_i)} \end{aligned}$$

- For the updates on the mixture component, extra care has to be given to ensure that they sum up to 1. We add a Lagrange multiplier λ with the condition $\sum_j P(j) - 1 = 0$

$$\frac{\partial \langle \mathcal{L} \rangle}{\partial P(j)} = 0 \Longleftrightarrow -\sum_{i=1}^{N} \frac{P_t(j|n_i)}{P(j)} + \lambda \frac{\partial}{\partial P(j)} \left(\sum_{k=1}^{M} P(i) - 1 \right) = 0$$
$$\Longleftrightarrow -\sum_{i=1}^{N} \frac{P_t(j|n_i)}{P(j)} + \lambda = 0$$
$$\Longleftrightarrow P(j) = -\frac{\sum_{i=1}^{N} P_t(j|n_i)}{\lambda}$$
$$\sum_{k=1}^{M} P(k) - 1 = 0$$
$$\Longleftrightarrow \sum_{k=1}^{M} -\frac{\sum_{i=1}^{N} P_t(k|n_i)}{\lambda} - 1 = 0$$
$$\Longleftrightarrow \lambda = -\sum_{k} \sum_{i=1}^{N} P_t(k|n_i)$$
$$\Longleftrightarrow \lambda = -N$$
$$\Longleftrightarrow P(j) = \frac{1}{N} \sum_{i=1}^{N} P_t(j|n_i)$$

- Combined E and M step:

$$P^*(j) = \frac{1}{N} \sum_{i=1}^{N} \frac{P(j) e^{-\theta_j} \theta_j^{n_i}}{\sum_{k=1}^{M} P(k) e^{-\theta_k} \theta_k^{n_i}}$$
$$\theta_j^* = \frac{1}{NP^*(j)} \sum_{i=1}^{N} \frac{n_i P(j) e^{-\theta_j} \theta_j^{n_i}}{\sum_{k=1}^{M} P(k) e^{-\theta_k} \theta_k^{n_i}}$$

(c) [CODE] Create a toy dataset with $N = 1000$ samples from a mixture of Poisson with $M = 3$, $\theta_1 = 1.0, \theta_2 = 20.0, \theta_3 = 50.0$ and $P(1) = P(2) = P(3) = 1/3$. Implement you EM algorithm to recover these parameters

```
mixpoisson (generic function with 1 method)
```

```julia
function mixpoisson(θ, p) # Return a mixture of Poissons with parameters theta and weights p
    MixtureModel(Poisson.(θ), p)
end
```

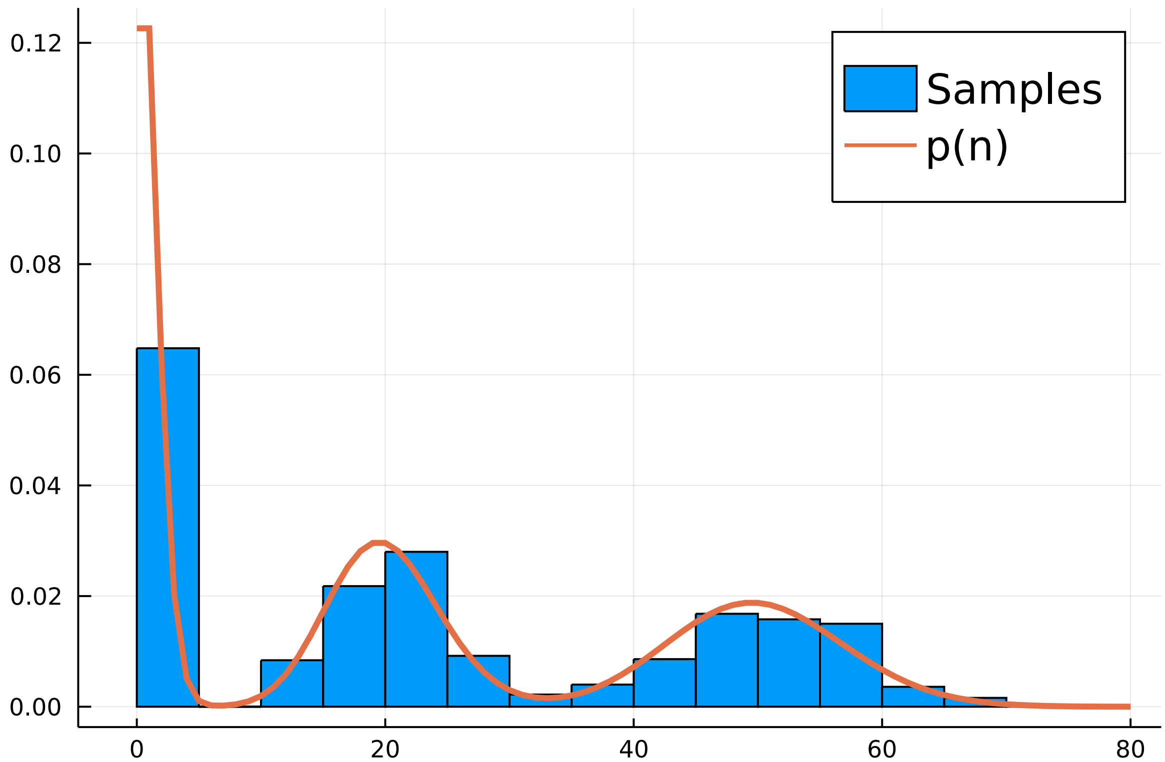
```
θ_true =  [1.0, 20.0, 50.0]; # Poisson parameters
```

```
p_true =  [1/3, 1/3, 1/3]; # Mixture parameters
```

```
d = mixpoisson(θ_true, p_true); # The true Poisson mixture
```

```
N = 1000; # Number of samples
```

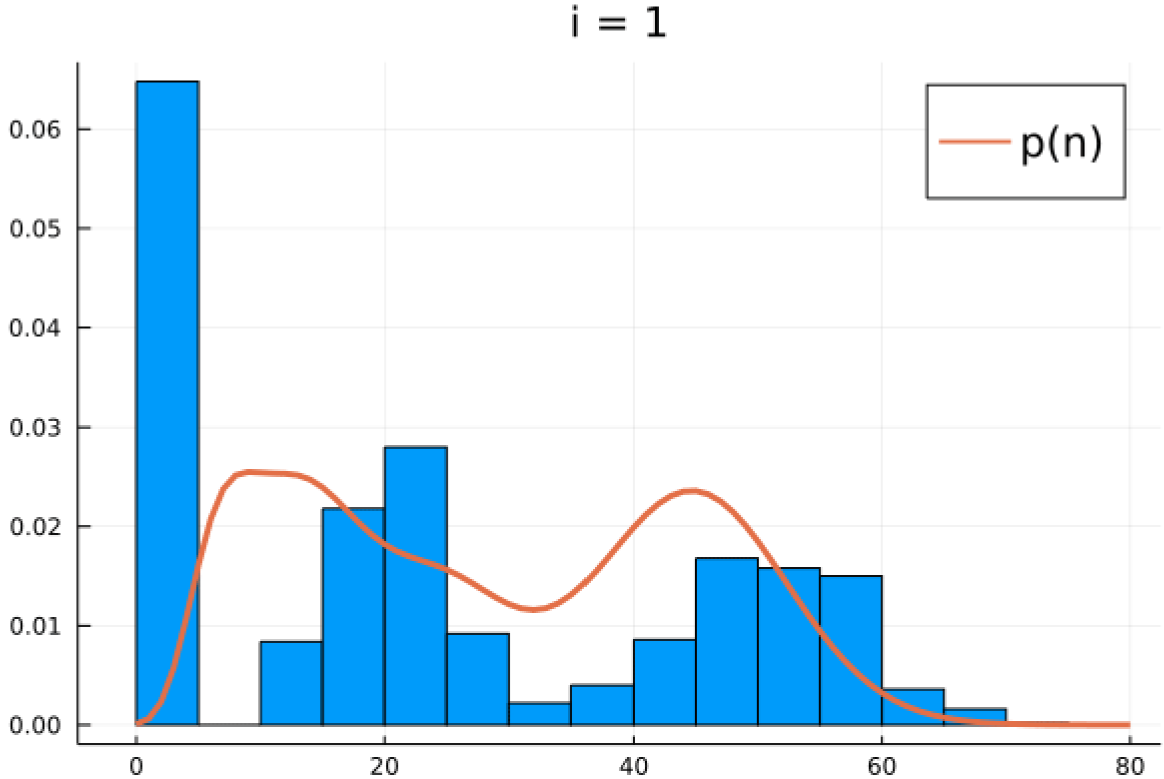
```
n = rand(d, N); # Sampled data
```



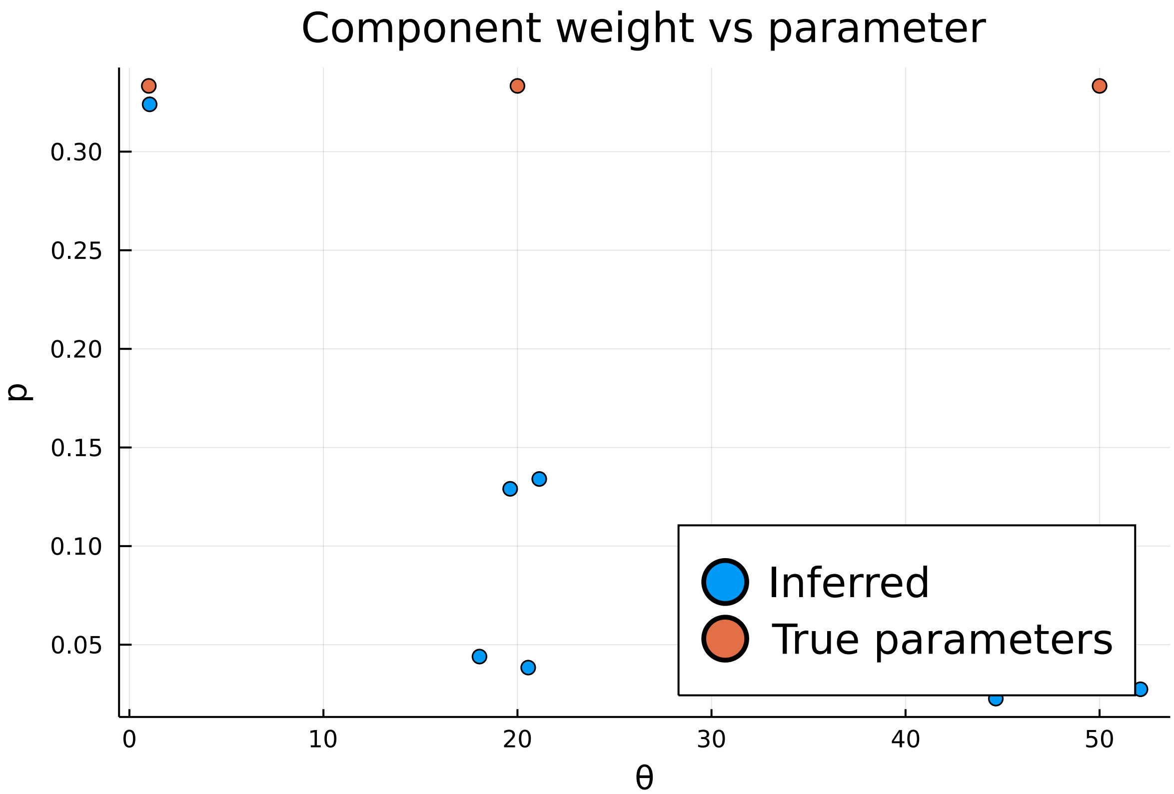

```
function pt(θ, p, n) # Compute Pt(p | θ, n)
    v = p .* exp.(-θ) .* θ .^ n
    v = v / sum(v)
end;
```

```
function update!(θ, p, n) # Update the parameters
    M = length(p)
    N = length(n)
    pvals = zeros(N, M)
    θvals = zeros(N, M)
    for i in 1:N # Loop over all the points
        x = pt(θ, p, n[i]) # Compute Pt for each j (x is a vector)
        pvals[i, :] = x # Save value
        θvals[i, :] = n[i] * x # Compute n * Pt
    end
    p .= vec(sum(pvals, dims = 1)) / N # Sum over the 1st dimension and take the mean
    θ .= vec(sum(θvals, dims = 1)) ./ vec(sum(pvals, dims = 1))
end;
```

Number of components 11




```
begin
    nIter = 10 # Number of iterations
    θ = rand(M) * 50 # Random initialization of the pararameters
    p = rand(M); p /= sum(p) # Random initialization of the weights and normalization
    anim = Animation() # Create an animation
    anim = @animate for i in 1:nIter # Run the algorithm for a few iterations
        d = mixpoisson(θ, p)
        histogram(n, nbins=20, normalize = true, lab = "", lw = 1.0)
        plot!(0:1:80,x->pdf(d, x), lab = "p(n)", title = "i = $(i)")
        update!(θ, p, n)
    end
    gif(anim, fps = 3)
end
```



2. Bayesian estimation for the Poisson distribution

Consider again the Poisson distribution for an integer valued random variable $n \in \{0, 1, 2, \dots\}$

$$P(n|\theta) = e^{-\theta} \frac{\theta^n}{n!} \, ,$$

- **<u>MATH</u> Write the Poisson distribution in the exponential family form :**

$$P(n|\theta) = f(n) \exp\left[\psi(\theta)\phi(n) + g(\theta)\right]$$

Solution

Writing

$$P(n|\theta) = \frac{1}{n!} e^{n \ln \theta - \theta}$$

wee see that $f(n) = \frac{1}{n!}$, $\phi(n) = n$, $\psi(\theta) = \ln\theta$ and $g(\theta) = -\theta$.

- **(b) [MATH] Use this exponential family representation to show that the conjugate prior for the Poisson distribution is given by the Gamma density**

$$p(\theta|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$$

where α, β are hyperparameters.

Solution

Following the lecture, the conjugate prior is of the form

$$p(\theta) \propto \exp\left[\psi(\theta)a + bg(\theta)\right] = \theta^a e^{-b\theta}$$

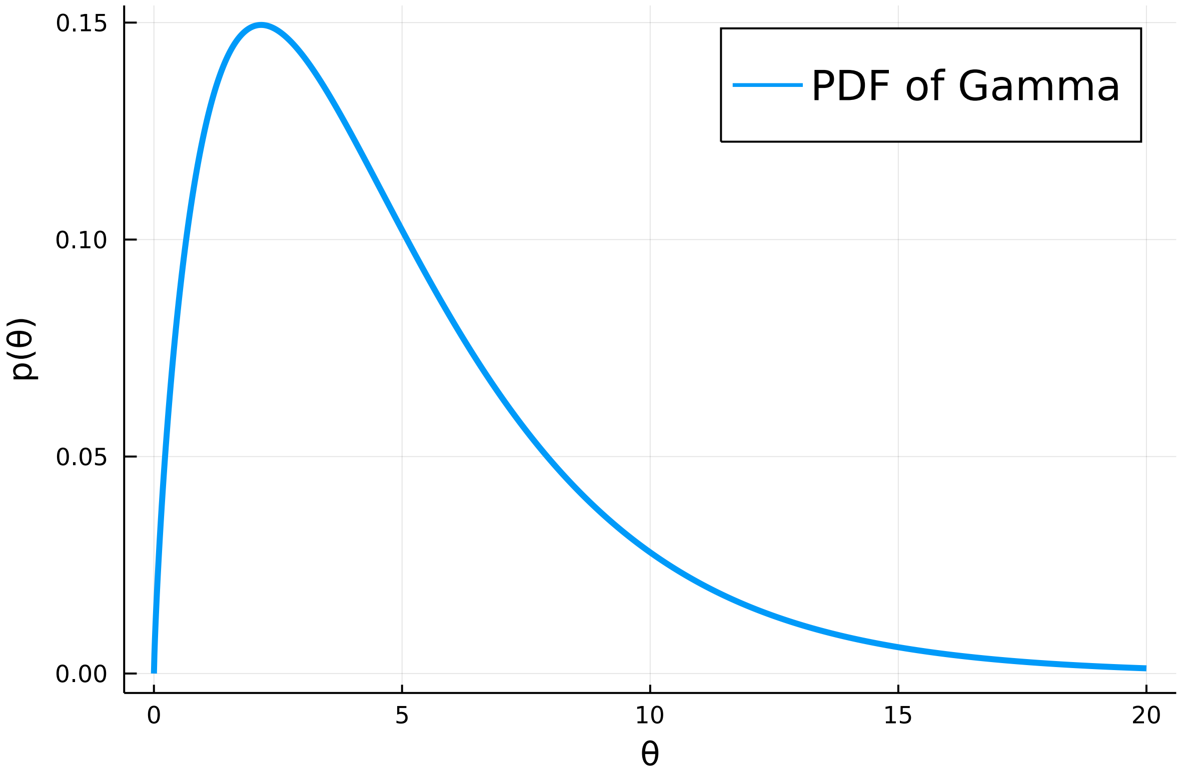
for some constants a, b. To make the density normalisable, we need $a > -1$ and $\beta > 0$. Setting $\beta \equiv b$ and $\alpha = b + 1$ we get the Gamma density. For the normalisation, we have

$$\int_0^\infty \theta^{\alpha-1}e^{-\beta\theta}\,d\theta = \beta^{-\alpha}\int_0^\infty y^{\alpha-1}e^{-y}\,dy = \beta^{-\alpha}\Gamma(\alpha)$$

where the last integral gives $\Gamma(\alpha)$, the {\em Euler Gamma–function}.

α_gamma = 1.8

β_gamma = 2.7



- **(c) [MATH] Assume that we observe Poisson data $D = (n_1, n_2, \ldots, n_N)$. Write down the posterior distribution $p(\theta|D)$ assuming the Gamma prior. What are the posterior mean and MAP estimators for θ ?**

Solution

The posterior distribution for θ is given by

$$p(\theta|D) = \frac{P(D|\theta)p(\theta|\alpha,\beta)}{P(D|\alpha,\beta)} \propto \prod_{i=1}^{N}\left(\theta^{n_i}e^{-\theta}\right)\,\theta^{\alpha-1}e^{-\beta\theta} = \theta^{\sum_{i=1}^{N}n_i+\alpha-1}e^{-(N+\beta)\theta}$$

This is again of the **Gamma form** with parameters $\beta' \doteq N + \beta$ and $\alpha' \doteq \sum_{i=1}^{N} n_i + \alpha$.

The MAP estimator is the one that maximises the exponent $-\beta'\theta + (\alpha' - 1)\ln\theta$ in the posterior. Taking the derivative wrt θ yields

$$\theta_{MAP} = \frac{\alpha' - 1}{\beta'} = \frac{\sum_{i=1}^{N} n_i + \alpha - 1}{N + \beta}$$

The posterior mean is defined by

$$\theta_{mean} = \int_0^\infty \theta \; p(\theta|D) = \frac{\beta'^{\alpha'}}{\Gamma(\alpha')} \int_0^\infty \theta^{\alpha'} e^{-\beta'\theta} \, d\theta = \frac{\beta'^{\alpha'}}{\Gamma(\alpha')} \frac{\Gamma(\alpha'+1)}{\beta'^{\alpha'+1}} = \frac{\alpha'}{\beta'}$$

In the last step we have used the relation $\Gamma(x+1) = x\Gamma(x)$.

- ## (d) [MATH] Compute the posterior variance for large N and compare your result with the asymptotic frequentist error of the maximum likelihood estimator.

Hint: For the computation of the frequentist error use the **Fisher Information** $J(\theta) \doteq E[(\frac{d \ln P(n|\theta)}{d\theta})^2]$ where the expectation is over the probability distribution $P(n|\theta)$.

Solution

The variance of the Gamma distribution given by :

$$\begin{aligned}\mathrm{Var}_{p(\theta|D)}(\theta) &= \frac{\alpha'}{\beta'^2} = \frac{\sum_{i=1}^N n_i + \alpha - 1}{(N+\beta)^2} \\ \lim_{N\to\infty} \mathrm{Var}_{p(\theta|D)}(\theta) &= 0\end{aligned}$$

$$\begin{aligned}J(\theta) =& E[(\frac{\partial C - \theta + n\log\theta}{\partial\theta})^2] \\ =& E[(-1 + n/\theta)^2] \\ =& E[1 - 2n/\theta + n^2/\theta^2] \\ =& 1 - 2 + 1 + 1/\theta \\ =& 1/\theta\end{aligned}$$

The error is then estimated by $J^{-1}(\theta)/N$ for $N \leftarrow \infty$. Here : $\frac{\theta}{N}$, which obviously converges to 0.

- ## (e) [CODE] Estimate the posterior distribution by continuously sampling from a Poisson distribution

and compare with the Maximum likelihood estimator.

θ_🐟= 10.0 : True Poisson parameter

```
d_poisson = Poisson(θ_poisson); # True Poisson distribution
```

```
alpha(n, α) = sum(n) + α; # Posterior of α
```

```
beta(N, β) = N + β; # Posterior for β
```

```
mapestimator(n, α, β) = (alpha(n, α) - 1) / beta(length(n), β);
```

```
mlestimator(n) = sum(n) / length(n);
```

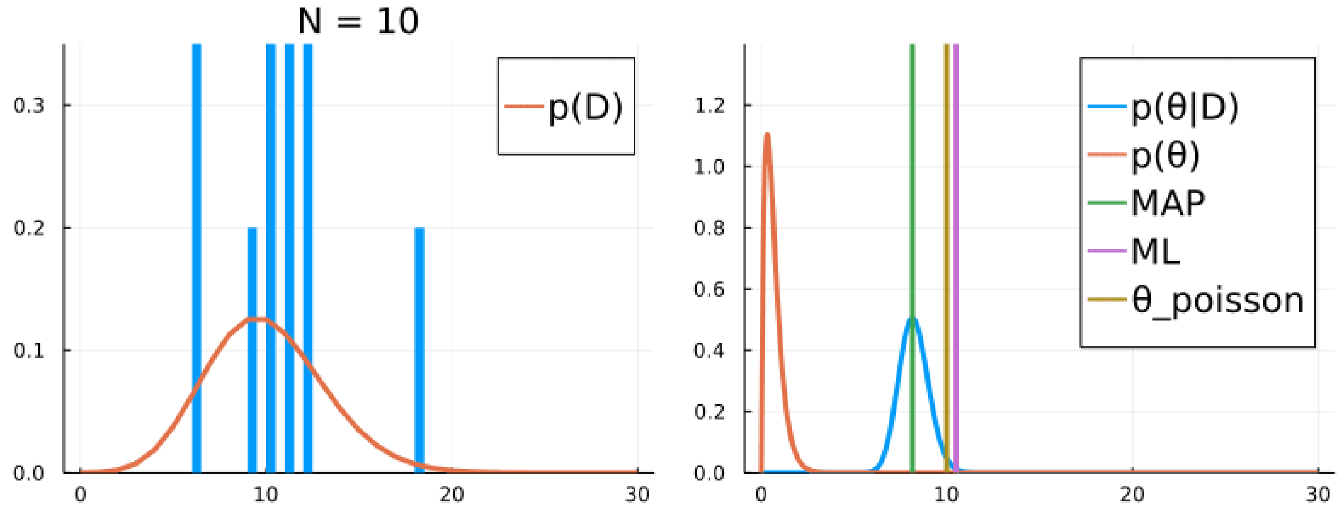
α = 2.0

β = 3.0

```
d_prior = Gamma(α, 1/β); # Prior distribution
```

```
begin # Elements for plotting
    nrange = 0:1:30
    xrange = 0:0.01:30
    Nmax = 50
    n_samples_per_step = 10
end;
```

```
begin
    n_model = Int[]
    anim_2 = @animate for i in 1:Nmax
        for _ in 1:n_samples_per_step
            push!(n_model, rand(d_poisson)) # Add n new samples
        end
        p1 = histogram(n_model; nbins=length(nrange), normalize=true, linewidth=0.0, title="N = $(i *  n_samples_per_step)", label="")
        plot!(nrange, x -> pdf(d_poisson, x), label="p(D)", ylims=(0, 0.35))
        d_posterior = Gamma(alpha(n_model, α),  1 / beta(length(n_model), β)) # Distributions.jl uses a different parametrization
        p2 = plot(xrange, x -> pdf(d_posterior, x), label="p(θ|D)")
        plot!(xrange, x -> pdf(d_prior, x); label="p(θ)")
        vline!([mapestimator(n_model, α, β)]; label="MAP", ylims=(0, 1.4))
        vline!([mlestimator(n_model)]; label="ML")
        vline!([θ_poisson]; label="θ_poisson")
        plot(p1, p2; size=(800, 300))
    end
end;
```




```
gif(anim_2, fps = 5)
```