Chapter 7. Statistical Estimation

7.5: Maximum A Posteriori Estimation

Slides (Google Drive) Alex Tsun Video (YouTube)

We've seen two ways now to estimate unknown parameters of a distribution. Maximum likelihood estimation (MLE) says that we should find the parameter θ that maximizes the likelihood ("probability") of seeing the data, whereas the method of moments (MoM) says that we should match as many moments as possible (mean, variance, etc.). Now, we learn yet another (and final) technique for estimation that will cover (there are many more...).

7.5.1 Maximum A Posteriori (MAP) Estimation

Maximum a Posteriori (MAP) estimation is quite different from the estimation techniques we learned so far (MLE/MoM), because it allows us to **incorporate prior knowledge** into our estimate. Suppose you wanted to estimate the unknown probability of heads on a coin θ: using MLE, you may flip the head 20 times and observe 13 heads, giving an estimate of 13/20. But what if your friend had flipped the coin before and observed 10 heads and 2 tails: how can you (formally) incorporate her information into your estimate? Or what if you just believed in general that coins were more likely to be fair $\theta = 0.5$ than unfair? We'll see how to do this below!

7.5.1.1 Intuition

In Maximum Likelihood Estimation (MLE), we used iid samples $\mathbf{x} = (x_1, \dots, x_n)$ from some distribution with unknown parameter(s) θ, in order to estimate θ.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \; L(\mathbf{x} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

Note: Recall that, using the English description, how we found $\hat{\theta}_{MLE}$ is: we computed this likelihood, which is the probability of seeing the data given the parameter θ, and we chose the "best" θ that maximized this likelihood.

You might have been thinking: shouldn't we be trying to maximize "$\mathbb{P}(\theta \mid x)$" instead? Well, this doesn't make sense **unless Θ is a R.V.**! And this is where Maximum A Posteriori (MAP) Estimation comes in.

So far, for MLE and MoM estimation, we assumed θ was fixed but unknown. This is called the **Frequentist framework** where we only estimate our parameter based on **data alone**, and θ is not a random variable. Now, we are in the **Bayesian framework**, meaning that our unknown parameter is a random variable Θ. This means, we will have some belief distribution $\pi_\Theta(\theta)$ (think of this as a density function over all possible values of the parameter), and after observing data $\mathbf{x}$, we will have a new/updated belief distribution $\pi_\Theta(\theta \mid \mathbf{x})$. Let's see a picture of what MAP is going to do first, before getting more into the math and formalism.

Example(s)

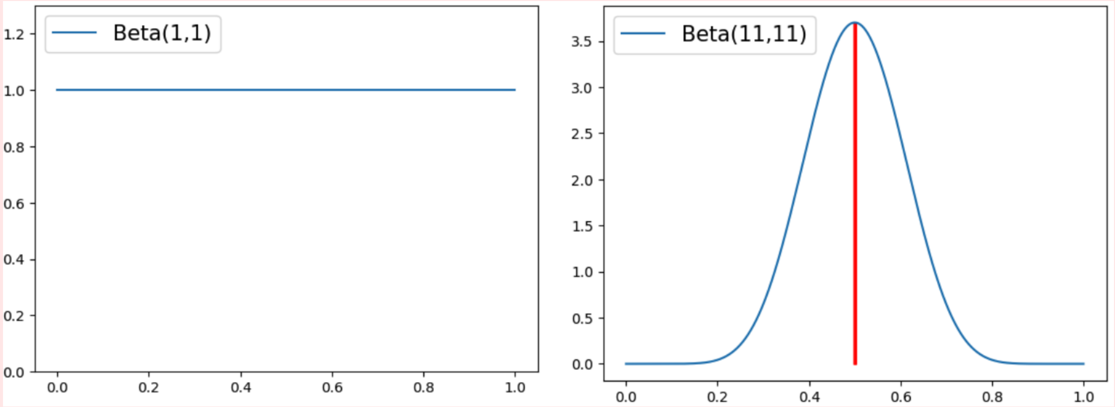
We'll see the idea of MAP being applied to our typical coin example. Suppose we are trying to estimate the unknown parameter for the probability of heads on a coin: that is, θ in $\text{Ber}(\theta)$. We are going to treat the parameter as a *random variable* (before in MLE/MoM we treated it as a *fixed* unknown quantity), so we'll call it Θ (capitalized θ).

1. **We must have a prior belief distribution $\pi_\Theta(\theta)$ over possible values that Θ could take on.**

   The range of Θ in our case is $\Omega_\Theta = [0, 1]$, because the probability of heads must be in this interval. Hence, when we plot the density function of Θ, the x-axis will range from 0 to 1.

   On a piece of paper, please sketch a density function that you might have for this probability of heads without yet seeing any data (coin flips). There are two reasonable shapes for this PDF:

   - The $\text{Unif}(0, 1) = \text{Beta}(1, 1)$ distribution (left picture below).
   - Some Beta distribution where $\alpha = \beta$, since most coins in this world are fair. Let's say $\text{Beta}(11, 11)$; meaning we pretend we've seen 10 heads and 10 tails (right picture below).



2. **We will observe some iid samples $\mathbf{x} = (x_1, \ldots, x_n)$.**

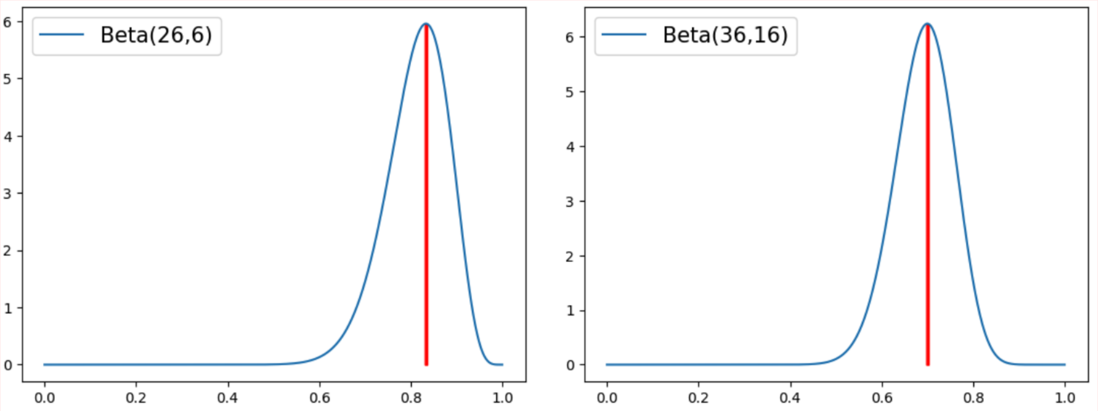
   Again, for the Bernoulli distribution, these will be a sequence of n 1's and 0's representing heads or tails. Suppose we observed $n = 30$ samples, in which $\sum_{i=1}^n x_i = 25$ were heads and $n - \sum_{i=1}^n x_i = 5$ were tails.

3. **We will combine our prior knowledge and the data to create a posterior belief distribution $\pi_\Theta(\theta \mid \mathbf{x})$.**

   Sketch two density functions for this posterior: one using the $\text{Beta}(1, 1)$ prior above, and one using the $\text{Beta}(11, 11)$ prior above. We'll compare these.

   - If our prior distribution was $\Theta \sim \text{Beta}(1, 1)$ (meaning we pretend we didn't see anything yet), then our posterior distribution should be $\Theta \mid \mathbf{x} \sim \text{Beta}(26, 6)$ (meaning we saw 25 heads and 5 tails total).

- If our prior distribution was $\Theta \sim \text{Beta}(11, 11)$ (meaning pretend we saw 10 heads and 10 tails beforehand), then our posterior distribution should be $\Theta \mid \mathbf{x} \sim \text{Beta}(36, 16)$ (meaning we saw 35 heads and 15 tails total).



4. **We'll give our MAP estimate as the *mode* of this posterior distribution. Hence, the name "Maximum a Posteriori".**

   - If we used the $\Theta \sim \text{Beta}(1, 1)$ prior, we ended up with the $\Theta \mid \mathbf{x} \sim \text{Beta}(26, 6)$ posterior, and our MAP estimate is defined to be the **mode** of the distribution, which occurs at $\hat{\theta}_{MAP} = \frac{25}{30} \approx 0.833$ (left picture above). You may notice that this would give the same as the MLE: we'll examine this more later!
   - If we used the $\Theta \sim \text{Beta}(11, 11)$ prior, we ended up with the $\Theta \mid \mathbf{x} \sim \text{Beta}(36, 16)$ posterior, our MAP estimate is defined to be the **mode** of the distribution, which occurs at $\hat{\theta}_{MAP} = \frac{35}{50} = 0.70$ (right picture above).

Hopefully you now see the process and idea behind MAP: We have a prior belief on our unknown parameter, and after observing data, we update our belief distribution and take the mode (most likely value)! Our estimate definitely depends on the prior distribution we choose (which is often arbitrary).

7.5.1.2 Derivation

We chose a Beta prior, and ended up with a Beta posterior, which made sense intuitively given our definition of the Beta distribution. But how do we prove this? We'll see the math behind MAP now (quite short), and see the same example again but mathematically rigorous now.

MAP Idea: Actually, unknown parameter(s) is a random variable Θ. We have a ***prior*** distribution (prior belief on Θ before seeing data) $\pi_\Theta(\theta)$ and ***posterior*** distribution (given data; updated belief on Θ after observing some data) $\pi_\Theta(\theta \mid \mathbf{x})$.

By Bayes' Theorem,

$$\pi_\Theta(\theta \mid \mathbf{x}) = \frac{L(\mathbf{x} \mid \theta)\pi_\Theta(\theta)}{\mathbb{P}(\mathbf{x})} \propto L(\mathbf{x} \mid \theta)\pi_\Theta(\theta)$$

Recall that π_Θ is just a PDF or PMF over possible values of Θ. In other words, now we are maximizing the ***posterior* distribution** $\pi_\Theta(\theta \mid x)$, where Θ has a PMF/PDF. That is, we are finding the *mode* of the density/mass function. Note that since the denominator $\mathbb{P}(x)$ in the expression above **does not** depend on θ, we can just maximize the numerator $L(\mathbf{x} \mid \theta)\pi_\Theta(\theta)$! Therefore:

$$\hat{\theta}_{MAP} = \arg\max_\theta\ \pi_\Theta(\theta \mid \mathbf{x}) = \arg\max_\theta\ L(\mathbf{x} \mid \theta)\pi_\Theta(\theta)$$

Definition 7.5.1: Maximum A Posteriori (MAP) Estimation

Let $x = (x_1, \dots, x_n)$ be iid realizations from probability mass function $p_X(t\,;\,\Theta = \theta)$ (if X discrete), or from density $f_X(t\,;\,\Theta = \theta)$ (if X continuous), where Θ is the random variable representing the parameter (or vector of parameters). We define the Maximum A Posteriori (MAP) estimator $\hat{\theta}_{MAP}$ of Θ to be the parameter which maximizes the **posterior** distribution of Θ given the data.

$$\hat{\theta}_{MAP} = \arg\max_\theta\ \pi_\Theta(\theta \mid \mathbf{x}) = \arg\max_\theta\ L(\mathbf{x} \mid \theta)\pi_\Theta(\theta)$$

That is, it's exactly the same as maximum likelihood, except instead of just maximizing the likelihood, we are maximizing the likelihood multiplied by the prior!

Now we'll see a similar coin-flipping example, but deriving the MAP estimate mathematically and building even more intuition. I encourage you to try each part out before reading the answers!

7.5.1.3 Example

Example(s)

(a) Suppose our samples are $\mathbf{x} = (0, 0, 1, 1, 0)$, from $\text{Ber}(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0, 1)$. What is the MLE for θ?

(b) Suppose we impose the restriction that $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

(c) Assume Θ is restricted as in part (b) (but now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1, \pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

(d) Show that we can make the MAP whatever we like, by finding a prior over $\{0.2, 0.5, 0.7\}$ so that the MAP is 0.2, another so that it is 0.5, and another so that it is 0.7.

(e) Typically, for the Bernoulli/Binomial distribution, if we use MAP, we want to be able to get any value $\in (0, 1)$, not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$. So we need a (continuous) prior distribution with range $(0, 1)$ instead of our discrete one. We assign $\Theta \sim \text{Beta}(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0, 1)$. Recall the mode of a $W \sim \text{Beta}(\alpha, \beta)$ random variable is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the mode is the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \dots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is $\frac{k}{n}$, where $k = \sum x_i$ (the total number of successes). Show that the posterior $\pi_\Theta(\theta \mid x)$ has a $\text{Beta}(k + \alpha, n - k + \beta)$ distribution, and find the MAP estimator.

(f) Recall that Beta$(1, 1) \equiv$ Unif$(0, 1)$ (pretend we saw $1-1$ heads and $1-1$ tails ahead of time). If we used this as the prior, how would the MLE and MAP compare?

(g) Since the posterior is also a Beta Distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution's parameter p. Interpret α, β as to how they affect our estimate. This is a really special property: if the prior distribution multipled by the likelihood results in a posterior distribution in the same family (with different parameters), then we say that distribution is the conjugate prior to the distribution we are estimating.

(h) As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our **prior** when n is small, or n is large?

(i) Which do you think is "better", MLE or MAP?

Solution

(a) Suppose our samples are $\mathbf{x} = (0, 0, 1, 1, 0)$, from Ber$(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0, 1)$. What is the MLE for θ?

- Answer: $\frac{2}{5}$. We just find the likelihood of the data, which is the probability of observing 2 heads and 3 tails, and find the θ that maximizes it.

$$L(\mathbf{x} \mid \theta) = \theta^2(1-\theta)^3$$
$$\hat{\theta}_{MLE} = \arg\max_{\theta \in [0,1]} \ \theta^2(1-\theta)^3 = \tfrac{2}{5}$$

(b) Suppose we impose the restriction that $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

- Answer: 0.5. We need to find which of the three acceptable θ values maximizes the likelihood, and since there are only finitely many, we can just plug them all in and compare!

$$L(\mathbf{x} \mid 0.2) = (0.2^2 0.8^3) = 0.02048$$
$$L(\mathbf{x} \mid 0.5) = (0.5^2 0.5^3) = 0.03125$$
$$L(\mathbf{x} \mid 0.7) = (0.7^2 0.3^3) = 0.01323$$
$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \{0.2, 0.5, 0.7\}} \ L(\mathbf{x} \mid \theta) = 0.5$$

(c) Assume Θ is restricted as in part (b) (but now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1, \pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

- Answer: 0.7. Instead of maximizing just the likelihood, we need to maximize the likelihood times the prior. Again, since there are only finitely many values, we just plug them in!

$$\pi_\Theta(0.2 \mid x) = L(\mathbf{x} \mid 0.2)\pi_\Theta(0.2) = (0.2^2 0.8^3)(0.1) = 0.0020480$$
$$\pi_\Theta(0.5 \mid x) = L(\mathbf{x} \mid 0.5)\pi_\Theta(0.5) = (0.5^2 0.5^3)(0.01) = 0.0003125$$
$$\pi_\Theta(0.7 \mid x) = L(\mathbf{x} \mid 0.7)\pi_\Theta(0.7) = (0.7^2 0.3^3)(0.89) = 0.0117747$$

Note the effect of this prior - by setting $\pi_\Theta(0.7)$ so high and the other two values, we actually get a different maximizer. This is the effect of the prior on the MAP estimate (which was completely arbitrary)!

(d) Show that we can make the MAP whatever we like, by finding a prior over $\{0.2, 0.5, 0.7\}$ so that the MAP is 0.2, another so that it is 0.5, and another so that it is 0.7.

- Answer: Choose $\pi_\Theta(\theta) = 1$ for the θ you want! This shows that the prior really does make a difference, and that MAP and MLE are indeed different techniques.

(e) Typically, for the Bernoulli/Binomial distribution, if we use MAP, we want to be able to get any value $\in (0,1)$, not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$. So we need a (continuous) prior distribution with range $(0,1)$ instead of our discrete one. We assign $\Theta \sim \text{Beta}(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$. Recall the mode of a $W \sim \text{Beta}(\alpha, \beta)$ random variable is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the mode is the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is $\frac{k}{n}$, where $k = \sum x_i$ (the total number of successes). Show that the posterior $\pi_\Theta(\theta \mid x)$ has a $\text{Beta}(k+\alpha, n-k+\beta)$ distribution, and find the MAP estimator.

- Answer: $\hat{\theta}_{MAP} = \frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}$. We first have to write out what the posterior distribution is, which is proportional to just the prior times the likelihood:

$$\pi_\Theta(\theta \mid x) \propto L(\mathbf{x} \mid \theta) \cdot \pi_\Theta(\theta)$$

$$= \left(\binom{n}{k}\theta^k(1-\theta)^{n-k}\right) \cdot \left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$\propto \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$$

The first to second line comes from noticing $L(\mathbf{x} \mid \theta)$ is just the probability of seeing exactly k successes out of n (binomial PMF), and plugging in our equation for π_Θ (beta density). The second to third line comes from dropping the normalizing constants (that don't depend on θ), which we can do because we only care to maximize this over θ. If you stare closely at that last equation, it actually proportional to the PDF of a Beta distribution with different parameters! Our posterior is hence $\text{Beta}(k+\alpha, n-k+\beta)$ since PDFs uniquely define a distribution (there is only one normalizing constant that would make it integrate to 1). The MAP estimator is the mode of this posterior Beta distribution, which is given by the formula:

$$\hat{\theta}_{MAP} = \frac{k+\alpha-1}{(k+\alpha-1)+(n-k+\beta-1)} = \frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}$$

Try staring at this to see why this might make sense. We'll explain it more in part (g)!

(f) Recall that $\text{Beta}(1,1) \equiv \text{Unif}(0,1)$ (pretend we saw $1-1$ heads and $1-1$ tails ahead of time). If we used this as the prior, how would the MLE and MAP compare?

- Answer: They would be the same! From our previous question, if $\alpha = \beta = 1$, then

$$\hat{\theta}_{MAP} = \frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)} = \frac{k}{n} = \hat{\theta}_{MLE}$$

This is because we don't have any prior information essentially, by saying each value is equally likely!

(g) Since the posterior is also a Beta Distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution's parameter p. Interpret α, β as to how they affect our estimate. This is a really special property: if the prior distribution multipled by the likelihood results in a posterior distribution in the same family (with different parameters), then we say that distribution is the conjugate prior to the distribution we are estimating.

- Answer: The interpretation is: pretend we saw $\alpha - 1$ heads ahead of time, and $\beta - 1$ tails ahead of time. Then our **total** number of heads is $k + (\alpha - 1)$ (real + fake) and our **total** number of trials is $n + (\alpha + \beta - 2)$ (real + fake), so that's our estimate! That's how prior information was factored in to our estimator, rather than just using what we actually saw in the data.

(h) As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our **prior** when n is small, or n is large?

- Answer: They become equal! The prior is important if we don't have much data, but as we get more, the evidence overwhelms the prior. You can imagine that if we only flipped the coin 5 times, the prior would play a huge role in our estimate. But if we flipped the coin 10,000 times, any (small) prior wouldn't really change our estimate.

(i) Which do you think is "better", MLE or MAP?

- Answer: There is no right answer. There are two main schools in statistics: Bayesians and Frequentists.
- Frequentists prefer MLE since they don't believe you should be putting a prior belief on anything, and you should only make judgment based on what you've seen. They believe the parameter being estimated is a **fixed quantity**.
- On the other hand, Bayesians prefer MAP, since they can incorporate their prior knowledge into the estimation. Hence the parameter being estimated is a **random variable**, and we seek the mode - the value with the highest probability or density. An example would be estimating the probability of heads of a coin - is it reasonable to assume it is more likely fair than not? If so, what distribution should we put on the parameter space?
- Anyway, in the long run, the prior "washes out", and the only thing that matters is the likelihood; the observed data. For small sample sizes like this, the prior significantly influences the MAP estimate. However, as the number of samples goes to infinity, the MAP and MLE are equal.

□

7.5.2 Exercises

1. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be iid samples from $\text{Exp}(\Theta)$ where Θ is a random variable (not fixed). Note that the range of Θ should be $\Omega_\Theta = [0, \infty)$ (the average rate of events per unit time), so any prior we choose should have this range.

   (a) Using the prior $\Theta \sim \text{Gamma}(r, \lambda)$ (for some arbitrary but known parameters $r, \lambda > 0$), show that the posterior distribution $\Theta \mid \mathbf{x}$ also follows a Gamma distribution and identify its parameters (by computing $\pi_\Theta(\theta \mid \mathbf{x})$). Then, explain this sentence: "The Gamma distribution is the conjugate prior for the rate parameter of the Exponential distribution". Hint: This can be done in just a few lines!

(b) Now derive the MAP estimate for Θ. The mode of a Gamma(s,ν) distribution is $\frac{s-1}{\nu}$. Hint: This should be just one line using your answer to part (a).

(c) Explain how this MAP estimate differs from the MLE estimate (recall for the Exponential distribution it was just the inverse sample mean $\frac{n}{\sum_{i=1}^{n} x_i}$), and provide an interpretation of r and λ as to how they affect the estimate.

Solution:

(a) Remember that the posterior is proportional to likelihood times prior, and the density of $Y \sim$ Exp(θ) is $f_Y(y \mid \theta) = \theta e^{-\theta y}$:

$$\begin{aligned}
\pi_\Theta(\theta \mid \mathbf{x}) &\propto L(\mathbf{x} \mid \theta)\pi_\Theta(\theta) && \text{[def of posterior]} \\
&= \left(\prod_{i=1}^{n} \theta e^{-\theta x_i}\right) \cdot \frac{\lambda^r}{(r-1)!}\theta^{r-1}e^{-\lambda\theta} && \text{[def of Exp}(\theta)\text{ likelihood + Gamma}(r,\lambda)\text{ pdf]} \\
&\propto \theta^n e^{-\theta \sum x_i}\theta^{r-1}e^{-\lambda\theta} && \text{[algebra, drop constants]} \\
&= \theta^{(n+r)-1}e^{-(\lambda+\sum x_i)\theta}
\end{aligned}$$

Therefore $\Theta \mid \mathbf{x} \sim$ Gamma$(n+r, \lambda + \sum x_i)$, since the final line above is proportional to the pdf for the gamma distribution (minus normalizing constant).

It is the conjugate prior because, assuming a Gamma prior for the Exponential likelihood, we end up with a Gamma posterior. That is, the prior and posterior are in the same family of distributions (Gamma) with different parameters.

(b) Just citing the mode of a Gamma given, we get

$$\hat{\theta}_{MAP} = \frac{n+r-1}{\lambda + \sum x_i}$$

(c) We see how the estimate changes from the MLE of $\hat{\theta}_{MLE} = \frac{n}{\sum x_i}$: pretend we saw $r-1$ extra events over λ units of time. (Instead of waiting for n events, we waited for $n+r-1$, and instead of $\sum x_i$ as our total time, we now have $\lambda + \sum x_i$ units of time).