Chapter 4

Basic Principles of Parameter Estimation

In probabilistic modeling, we are typically presented with a set of observations and the objective is to find a model, or function, $\hat{f}$ that shows good agreement with the data and respects certain additional requirements. We shall roughly categorize these requirements into three groups: (*i*) the ability to generalize well, (*ii*) the ability to incorporate prior knowledge and assumptions into modeling, and (*iii*) scalability. First, the model should be able to stand the test of time; that is, its performance on the previously unseen data should not deteriorate once this new data is presented. Models with such performance are said to generalize well. Second, $\hat{f}$ must be able to incorporate information about the model space $\mathcal{F}$ from which it is selected and the process of selecting a model should be able to accept training "advice" from an analyst. Finally, when large amounts of data are available, learning algorithms must be able to provide solutions in reasonable time given the resources such as memory or CPU power. In summary, the choice of a model ultimately depends on the observations at hand, our experience with modeling real-life phenomena, and the ability of algorithms to find good solutions given limited resources.

An easy way to think about finding the "best" model is through learning parameters of a distribution. Suppose we are given a set of observations $\mathcal{D} = \{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}$ and have knowledge that $\mathcal{F}$ is a family of all univariate Gaussian distributions; e.g., $\mathcal{F} = \text{Gaussian}(\mu, \sigma^2)$, with $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$. In this case, the problem of finding the best model (by which we mean function) can be seen as finding the best parameters μ^* and σ^*; i.e., the problem can be seen as *parameter estimation.* We call this process estimation because the typical assumption is that the data was generated by an unknown model from $\mathcal{F}$ whose parameters we are trying to recover from data.

We will formalize parameter estimation using probabilistic techniques and will subsequently find solutions through optimization, occasionally with constrains in the parameter space. The main assumption throughout this part will be that the set of observations $\mathcal{D}$ was generated (or collected) independently and according to the same distribution $p(x)$. The statistical framework for model inference is shown in Figure 4.1.

4.1 Maximum a posteriori and maximum likelihood estimation

The idea behind *maximum a posteriori* (MAP) estimation is to find the most probable model for the observed data. Given the data set $\mathcal{D}$, we formalize the MAP solution as

$$f_{\text{MAP}} = \arg\max_{f \in \mathcal{F}} \{p(f|\mathcal{D})\},$$

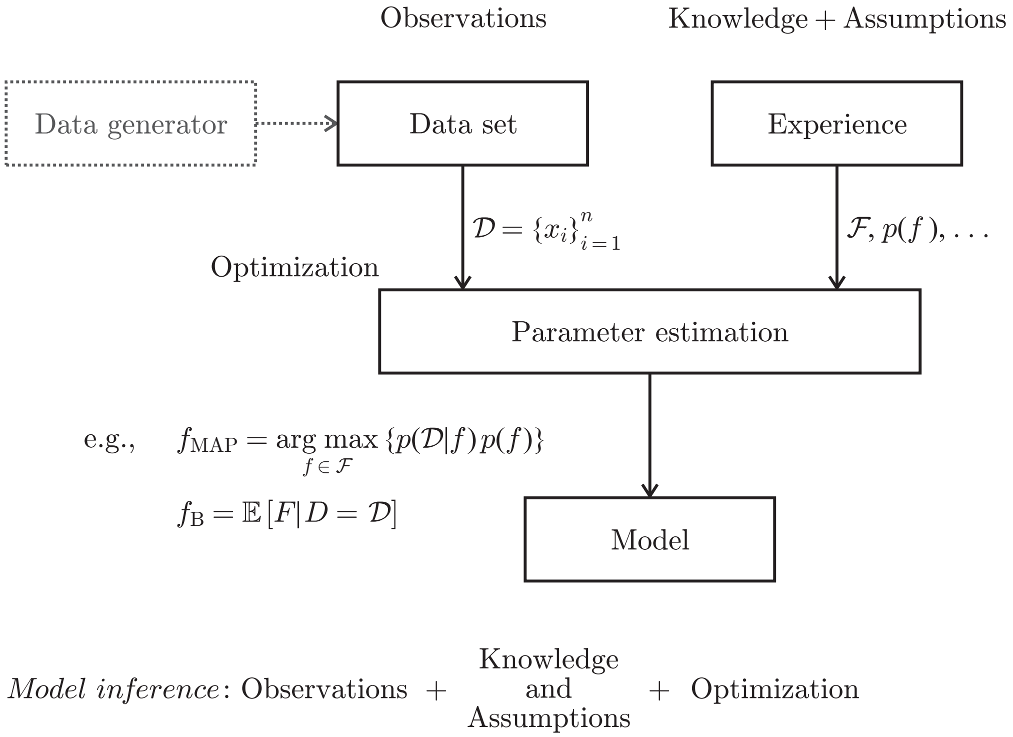
Observations
Knowledge + Assumptions
Data generator
Data set
Experience
$\mathcal{D} = \{x_i\}_{i=1}^{n}$
$\mathcal{F}, p(f), \ldots$
Optimization
Parameter estimation
e.g., $f_{\text{MAP}} = \arg\max_{f \in \mathcal{F}} \{p(\mathcal{D}|f)\, p(f)\}$
$f_{\text{B}} = \mathbb{E}\,[F|D = \mathcal{D}]$
Model
Model inference: Observations + Knowledge and Assumptions + Optimization

Figure 4.1: Statistical framework for model inference. The estimates of the parameters are made using a set of observations $\mathcal{D}$ as well as experience in the form of model space $\mathcal{F}$, prior distribution $p(f)$, or specific starting solutions in the optimization step.

where $p(f|\mathcal{D})$ is called the *posterior distribution* of the model given the data. In discrete model spaces, $p(f|\mathcal{D})$ is the probability mass function and the MAP estimate is exactly the most probable model. Its counterpart in continuous spaces is the model with the largest value of the posterior density function. Note that we use words *model*, which is a function, and its *parameters*, which are the coefficients of that function, somewhat interchangeably. However, we should keep in mind the difference, even if only for pedantic reasons.

To calculate the posterior distribution we start by applying the Bayes rule as

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})}, \tag{4.1}$$

where $p(\mathcal{D}|f)$ is called the *likelihood*function, $p(f)$ is the *prior* distribution of the model, and $p(\mathcal{D})$ is the *marginal* distribution of the data. Notice that we use $\mathcal{D}$ for the observed data set, but that we usually think of it as a realization of a multidimensional random variable D drawn according to some distribution $p(\mathcal{D})$. We can use marginalization to express $p(\mathcal{D})$ as

$$p(\mathcal{D}) = \begin{cases} \sum_{f \in \mathcal{F}} p(\mathcal{D}|f)p(f) & f : \text{discrete} \\ \\ \int_{\mathcal{F}} p(\mathcal{D}|f)p(f)df & f : \text{continuous} \end{cases}$$

Therefore, the posterior distribution can be fully described using the likelihood and the prior. The field of research and practice involving ways to determine this distribution and optimal models is referred to as *inferential statistics.* The posterior distribution is sometimes referred to as inverse probability.

Finding f_{MAP} can be greatly simplified because $p(\mathcal{D})$ in the denominator does not affect maximization. We shall re-write Equation (4.1) as

$$\begin{aligned} p(f|\mathcal{D}) &= \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|f) \cdot p(f), \end{aligned}$$

where $\propto$ is the proportionality symbol. Thus, we can find the MAP solution by solving the following optimization problem

$$f_{\text{MAP}} = \arg\max_{f \in \mathcal{F}} \left\{ p(\mathcal{D}|f) p(f) \right\}.$$

In some situations we may not have a reason to prefer one model over another and can think of $p(f)$ as a constant over the model space $\mathcal{F}$. Then, maximum a posteriori estimation reduces to the maximization of the likelihood function; i.e.,

$$f_{\text{ML}} = \arg\max_{f \in \mathcal{F}} \left\{ p(\mathcal{D}|f) \right\}.$$

We will refer to this solution as the *maximum likelihood* solution. Formally speaking, the assumption that $p(f)$ is constant is problematic because a uniform distribution cannot be always defined (say, over $\mathbb{R}$), though there are some solutions to this issue using improper priors. Nonetheless, it may be useful to think of the maximum likelihood approach as a separate technique, rather than a special case of MAP estimation, but keep this connection in mind.

Observe that MAP and ML approaches report solutions corresponding to the mode of the posterior distribution and the likelihood function, respectively. We shall later contrast this estimation technique with the view of the Bayesian statistics in which the goal is to minimize the posterior risk. Such estimation typically results in calculating conditional expectations, which can be complex integration problems. From a different point of view, MAP and ML estimates are called *point estimates*, as opposed to estimates that report confidence intervals for a particular group of parameters.

Example 9: Suppose data set $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ is an i.i.d. sample from a Poisson distribution with a fixed but unknown parameter λ_0. Find a maximum likelihood estimate of λ_0.

The probability mass function of a Poisson distribution is expressed as $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$, with some parameter $\lambda \in \mathbb{R}^+$. We will estimate this parameter as

$$\lambda_{\text{ML}} = \arg\max_{\lambda \in (0,\infty)} \left\{ p(\mathcal{D}|\lambda) \right\}. \tag{4.2}$$

We can write the likelihood function as

$$\begin{aligned} p(\mathcal{D}|\lambda) &= p(\{x_i\}_{i=1}^n |\lambda) \\ &= \prod_{i=1}^n p(x_i|\lambda) \\ &= \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}. \end{aligned}$$

To find λ that maximizes the likelihood, we will first take a logarithm (a monotonic function) to simplify the calculation, then find its first derivative with respect to λ, and finally equate it with zero to find the maximum. Specifically, we express the log-likelihood $ll(D, \lambda) = \ln p(\mathcal{D}|\lambda)$ as

$$ll(\mathcal{D}, \lambda) = \ln \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln (x_i!)$$

and proceed with the first derivative as

$$\begin{aligned} \frac{\partial ll(\mathcal{D}, \lambda)}{\partial \lambda} &= \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n \\ &= 0. \end{aligned}$$

By substituting $n = 6$ and values from $\mathcal{D}$, we can compute the solution as

$$\begin{aligned} \lambda_{\mathrm{ML}} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ &= 5.5, \end{aligned}$$

which is simply a sample mean. The second derivative of the likelihood function is always negative because λ must be positive; thus, the previous expression indeed maximizes the likelihood. Note that to properly maximize this loss, we also need to ensure the constraint $\lambda \in (0, \infty)$ is enforced. Because the solution above is in the constraint set, we know we have the correct solution to Equation (4.2); however, in other situations, we will have to explicitly enforce constraints in the optimization, as we will discuss later.

□

Example 10: Let $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ again be an i.i.d. sample from Poisson(λ_0), but now we are also given additional information. Suppose the prior knowledge about λ_0 can be expressed using a gamma distribution $\Gamma(x|k, \theta)$ with parameters $k = 3$ and $\theta = 1$. Find the maximum a posteriori estimate of λ_0.

First, we write the probability density function of the gamma family as

$$\Gamma(x|k, \theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)},$$

where $x > 0$, $k > 0$, and $\theta > 0$. $\Gamma(k)$ is the gamma function that generalizes the factorial function; when k is an integer, we have $\Gamma(k) = (k-1)!$. The MAP estimate of the parameters can be found as

$$\lambda_{\mathrm{MAP}} = \underset{\lambda \in (0,\infty)}{\arg\max} \left\{ p(\mathcal{D}|\lambda) p(\lambda) \right\}.$$

As before, we can write the likelihood function as

$$p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

and the prior distribution as

$$p(\lambda) = \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}.$$

Now, we can maximize the logarithm of the posterior distribution $p(\lambda|\mathcal{D})$ using

$$\begin{aligned}\ln p(\lambda|\mathcal{D}) &\propto \ln p(\mathcal{D}|\lambda) + \ln p(\lambda) \\ &= \ln \lambda(k - 1 + \sum_{i=1}^{n} x_i) - \lambda(n + \frac{1}{\theta}) - \sum_{i=1}^{n} \ln x_i! - k \ln \theta - \ln \Gamma(k)\end{aligned}$$

to obtain

$$\begin{aligned}\lambda_{\text{MAP}} &= \frac{k - 1 + \sum_{i=1}^{n} x_i}{n + \frac{1}{\theta}} \\ &= 5\end{aligned}$$

after incorporating all data.

A quick look at λ_{MAP} and λ_{ML} suggests that as n grows, both numerators and denominators in the expressions above become increasingly more similar. In fact, it is a well-known result that, in the limit of infinite samples, both the MAP and ML converge to the same model, f, as long as the prior does not have zero probability (or density) on f. This result shows that the MAP estimate approaches the ML solution for large data sets. In other words, large data diminishes the importance of prior knowledge. This is an important conclusion because it simplifies mathematical apparatus necessary for practical inference.

To get some intuition for this result, we will show that the MAP and ML estimates converge to the same solution for the above example with a Poisson distribution. Let $s_n = \sum_{i=1}^{n} x_i$, which is a sample from the random variable $S_n = \sum_{i=1}^{n} X_i$. If $\lim_{n\to\infty} s_n/n^2 = 0$ (i.e., s_n does not grow faster than n^2), then

$$\begin{aligned}|\lambda_{\text{MAP}} - \lambda_{\text{ML}}| &= \left| \frac{k - 1 + s_n}{n + 1/\theta} - \frac{s_n}{n} \right| \\ &= \left| \frac{k - 1}{n + 1/\theta} - \frac{s_n}{n(n + 1/\theta)} \right| \\ &\leq \frac{|k - 1|}{n + 1/\theta} + \frac{s_n}{n(n + 1/\theta)} \qquad \xrightarrow[n\to\infty]{} 0\end{aligned}$$

Note that if $\lim_{n\to\infty} s_n/n^2 \neq 0$, then both estimators go to ∞; however, such a sequence of values has an essentially zero probability of occurring. Consistency theorems for ML and MAP estimation state that convergence to the true parameters occurs "almost surely" or "with probability 1" to indicate that these unbounded sequences constitute a set of measure-zero, under certain reasonable conditions. More details can be found in Larry Wasserman's book "All of Statistics" (Theorem 9.13).

□

Example 11: Let $\mathcal{D} = \{x_i\}_{i=1}^{n}$ be an i.i.d. sample from a univariate Gaussian distribution. Find the maximum likelihood estimates of the parameters.

We start by forming the log-likelihood function

$$\begin{aligned}\ln p(\mathcal{D}|\mu,\sigma) &= \ln \prod_{i=1}^{n} p(x_i|\mu,\sigma) \\ &= n \ln \frac{1}{\sqrt{2\pi}} + n \ln \frac{1}{\sigma} - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}.\end{aligned}$$

We now compute the partial derivatives of the log-likelihood with respect to all parameters as

$$\frac{\partial}{\partial \mu} \ln p(\mathcal{D}|\mu,\sigma) = \frac{\sum_{i=1}^{n} (x_i - \mu)}{\sigma^2}$$

and

$$\frac{\partial}{\partial \sigma} \ln p(\mathcal{D}|\mu,\sigma) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^3}.$$

From here, we can proceed to derive that

$$\mu_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and

$$\sigma^2_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{\mathrm{ML}})^2.$$

□

4.2 Maximum likelihood for conditional distributions

We can also formulate maximum likelihood problems for conditional distributions. Recall that a conditional distribution has the form $p(y|x)$, for two random variables Y and X, where above we considered the marginal distribution $p(x)$ or $p(y)$. For the distributions above, we asked: what is the distribution over this variable? For a conditional distribution, we are instead asking: given some auxiliary information, now what is the distribution over this variable? When the auxiliary information changes, so will the distribution over the variable. For example, we may want to condition a distribution over sales of a particular product (Y) given the current month (X). We expect the distribution over Y to be different, depending on the month.

Conditional distributions can be from any of the distribution families discussed above, and we can similarly formulate parameter estimation problems. The parameters, however, are usually tied to the given variable X. We provide a simple example to demonstrate this below. Much of the parameter estimation formulations we consider later will be for conditional distributions, because in machine learning we typically have a large number of auxiliary variables (features) and are trying to predict (or learn the distribution over) targets.

Example 12: Assume we are given two random variables X and Y and that you believe $p(y|x) = \mathcal{N}(\mu = x, \sigma^2)$ for some unknown σ. Our goal is to estimate this unknown parameter

σ. Notice that the distribution over Y varies, depending on which X value is observed or given.

We again start by forming the log-likelihood function, now for a set pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, and apply the chain rule, $p(x_i, y_i) = p(y_i|x_i)p(x_i)$, to solve this problem.

$$\begin{aligned}
\ln p(\mathcal{D}|\sigma) &= \ln \prod_{i=1}^n p(x_i, y_i|\sigma) \\
&= \ln \prod_{i=1}^n p(y_i|x_i, \sigma)p(x_i) \\
&= n \ln \frac{1}{\sqrt{2\pi}} + n \ln \frac{1}{\sigma} - \frac{\sum_{i=1}^n (y_i - x_i)^2}{2\sigma^2} + \sum_{i=1}^n \ln p(x_i).
\end{aligned}$$

Observe that the middle line above incorporates that x_i is independent of σ, and that the last line uses $\mu = x_i$ for each normal distribution $p(y_i|x_i, \sigma)$. We now compute the derivative of the log-likelihood with respect to the parameter σ as

$$\frac{\partial}{\partial \sigma} \ln p(\mathcal{D}|\sigma) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (y_i - x_i)^2}{\sigma^3},$$

where $\frac{\partial}{\partial \sigma} \sum_{i=1}^n \ln p(x_i) = 0$ because σ does not parameterize $p(x_i)$. Therefore, to obtain the optimal σ, we do not need to know or specify the distribution over the random variable X. By setting the derivative to zero, to obtain a stationary point, we obtain

$$\sigma^2_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2.$$

□

4.2.1 The relationship with Kullback-Leibler divergence

We now investigate the relationship between maximum likelihood estimation and Kullback-Leibler divergence. Kullback-Leibler divergence between two probability distributions $p(x)$ and $q(x)$ is defined on $\mathcal{X} = \mathbb{R}$ as

$$D_{\mathrm{KL}}(p||q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx.$$

In information theory, Kullback-Leibler divergence has a natural interpretation of the inefficiency of signal compression when the code is constructed using a suboptimal distribution $q(x)$ instead of the correct (but unknown) distribution $p(x)$ according to which the data has been generated. However, more often than not, Kullback-Leibler divergence is simply considered to be a measure of divergence between two probability distributions. Although this divergence is not a metric (it is not symmetric and does not satisfy the triangle inequality) it has important theoretical properties in that (*i*) it is always non-negative and (*ii*) it is equal to zero if and only if $p(x) = q(x)$.

Consider now a divergence between an estimated probability distribution $p(x|\theta)$ and an underlying (true) distribution $p(x|\theta_0)$ according to which the data set $\mathcal{D} = \{x_i\}_{i=1}^n$ was generated. The Kullback-Leibler divergence between $p(x|\theta)$ and $p(x|\theta_0)$ is

$$\begin{aligned} D_{\mathrm{KL}}(p(x|\theta_0)||p(x|\theta)) &= \int_{-\infty}^{\infty} p(x|\theta_0)\log\frac{p(x|\theta_0)}{p(x|\theta)}dx \\ &= \int_{-\infty}^{\infty} p(x|\theta_0)\log\frac{1}{p(x|\theta)}dx - \int_{-\infty}^{\infty} p(x|\theta_0)\log\frac{1}{p(x|\theta_0)}dx. \end{aligned}$$

The second term in the above equation is simply the (differential) entropy of the true distribution and is not influenced by our choice of the model θ. The first term, on the other hand, can be expressed as

$$\int_{-\infty}^{\infty} p(x|\theta_0)\log\frac{1}{p(x|\theta)}dx = -\mathbb{E}[\log p(X|\theta)]$$

Therefore, maximizing $\mathbb{E}[\log p(X|\theta)]$ minimizes the Kullback-Leibler divergence between $p(x|\theta)$ and $p(x|\theta_0)$. Using the strong law of large numbers, we know that

$$\frac{1}{n}\sum_{i=1}^{n}\log p(x_i|\theta) \quad \overset{a.s.}{\to} \quad \mathbb{E}[\log p(X|\theta)]$$

when $n \to \infty$. Thus, when the data set is sufficiently large, maximizing the likelihood function minimizes the Kullback-Leibler divergence and leads to the conclusion that $p(x|\theta_{\mathrm{ML}}) = p(x|\theta_0)$, if the underlying assumptions are satisfied. Under reasonable conditions, we can infer from it that $\theta_{\mathrm{ML}} = \theta_0$. This will hold for families of distributions for which a set of parameters uniquely determines the probability distribution; e.g., it will not generally hold for mixtures of distributions but we will discuss this situation later. The relationship between maximum likelihood estimation and minimizing Kullback-Leibler divergence is only one of the many connections between statistics and information theory.

4.3 Bayesian estimation

Maximum a posteriori and maximum likelihood approaches report the solution that corresponds to the mode of the posterior distribution and the likelihood function, respectively. This approach, however, does not effectively address the possibility of skewed distributions, multimodal distributions or simply large regions with similar values of $p(f|\mathcal{D})$. Bayesian estimation addresses those concerns.

The main idea in Bayesian statistics is minimization of the *posterior risk*

$$R = \int_{\mathcal{F}} \ell(f, \hat{f}) \cdot p(f|\mathcal{D})df,$$

where $\hat{f}$ is our estimate and $\ell(f, \hat{f})$ is some loss function between two models. When $\ell(f, \hat{f}) = (f - \hat{f})^2$ (ignore the abuse of notation), we can minimize the posterior risk as follows

$$\begin{aligned} \frac{\partial}{\partial \hat{f}}R &= 2\hat{f} - 2\int_{\mathcal{F}} f \cdot p(f|\mathcal{D})df \\ &= 0 \end{aligned}$$

from which it can be derived that the minimizer of the posterior risk is the posterior mean function; i.e.,

$$\begin{aligned} f_{\mathrm{B}} &= \int_{\mathcal{F}} f \cdot p(f|\mathcal{D}) df \\ &= \mathbb{E}[F|\mathcal{D}], \end{aligned}$$

where F is a random variable representing the model. We shall refer to f_{B} as the Bayes estimator. It is important to mention that computing the posterior mean usually involves solving complex integrals. In some situations, these integrals can be solved analytically; in others, numerical integration is necessary.

Example 13: Let $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ yet again be an i.i.d. sample from Poisson(λ_0). Suppose the prior knowledge about the parameter of the distribution can be expressed using a gamma distribution with parameters $k = 3$ and $\theta = 1$. Find the Bayesian estimate of λ_0.

We want to find $\mathbb{E}[\Lambda|\mathcal{D}]$. Let us first write the posterior distribution as

$$\begin{aligned} p(\lambda|\mathcal{D}) &= \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})} \\ &= \frac{p(\mathcal{D}|\lambda)p(\lambda)}{\int_0^\infty p(\mathcal{D}|\lambda)p(\lambda)d\lambda}, \end{aligned}$$

where, as shown in previous examples, we have that

$$p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}$$

and

$$p(\lambda) = \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}.$$

Before calculating $p(\mathcal{D})$, let us first note that

$$\int_0^\infty x^{\alpha-1} e^{-\beta x} dx = \frac{\Gamma(\alpha)}{\beta^\alpha}.$$

Now, we can derive that

$$\begin{aligned} p(\mathcal{D}) &= \int_0^\infty p(\mathcal{D}|\lambda)p(\lambda)d\lambda \\ &= \int_0^\infty \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!} \cdot \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)} d\lambda \\ &= \frac{\Gamma(k + \sum_{i=1}^n x_i)}{\theta^k \Gamma(k) \prod_{i=1}^n x_i! (n + \frac{1}{\theta})^{\sum_{i=1}^n x_i + k}} \end{aligned}$$

and subsequently that

$$
\begin{aligned}
p(\lambda|\mathcal{D}) &= \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})} \\
&= \frac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \cdot \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)} \cdot \frac{\theta^k \Gamma(k) \prod_{i=1}^{n} x_i!(n+\frac{1}{\theta})^{\sum_{i=1}^{n} x_i + k}}{\Gamma(k+\sum_{i=1}^{n} x_i)} \\
&= \frac{\lambda^{k-1+\sum_{i=1}^{n} x_i} \cdot e^{-\lambda(n+1/\theta)} \cdot (n+\frac{1}{\theta})^{\sum_{i=1}^{n} x_i + k}}{\Gamma(k+\sum_{i=1}^{n} x_i)}.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\mathbb{E}[\Lambda|\mathcal{D}] &= \int_0^\infty \lambda p(\lambda|\mathcal{D}) d\lambda \\
&= \frac{k+\sum_{i=1}^{n} x_i}{n+\frac{1}{\theta}} \\
&= 5.14
\end{aligned}
$$

which is nearly the same solution as the MAP estimate found in Example 9. □

It is evident from the previous example that selection of the prior distribution has important implications on calculation of the posterior mean. We have not picked the gamma distribution by chance; that is, when the likelihood was multiplied by the prior, the resulting distribution remained in the same class of functions as the prior. We shall refer to such prior distributions as *conjugate priors*. Conjugate priors are also simplifying the mathematics; in fact, this is a major reason for their consideration. Interestingly, in addition to the Poisson distribution, the gamma distribution is a conjugate prior to the exponential distribution as well as the gamma distribution itself.

4.4 Parameter estimation for mixtures of distributions

We now investigate parameter estimation for mixture models, which is most commonly carried out using the expectation-maximization (EM) algorithm. As before, we are given a set of i.i.d. observations $\mathcal{D} = \{x_i\}_{i=1}^n$, with the goal of estimating the parameters of the mixture distribution

$$
p(x|\theta) = \sum_{j=1}^{m} w_j p(x|\theta_j),
$$

where all coefficients w_j are nonnegative and sum to one. In the equation above, we used $\theta = (w_1, w_2, \ldots, w_m, \theta_1, \theta_2, \ldots, \theta_m)$ to combine all parameters. For now, we shall assume that m is given and will address simultaneous estimation of θ and m later.

Let us attempt to find the maximum likelihood solution first. By plugging the formula for $p(x|\theta)$ into the likelihood function we obtain

$$
\begin{aligned}
p(\mathcal{D}|\theta) &= \prod_{i=1}^{n} p(x_i|\theta) \\
&= \prod_{i=1}^{n} \left(\sum_{j=1}^{m} w_j p(x_i|\theta_j) \right),
\end{aligned}
\tag{4.3}
$$

which, unfortunately, is difficult to maximize using differential calculus (why?).[1] We therefore need a different approach.

4.4.1 Basic iterative estimation for mixtures of distributions

Before introducing the EM algorithm, let us for a moment present two hypothetical scenarios that will help us to understand the algorithm and the principles behind it. First, suppose that information is available as to which mixing component generated which data point. That is, suppose that $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ is an i.i.d. sample from some distribution $p(x, y)$, where $y \in \mathcal{Y} = \{1, 2, \ldots, m\}$ specifies the mixing component. How would the maximization be performed then? Let us write the likelihood function as

$$
\begin{aligned}
p(\mathcal{D}|\theta) &= \prod_{i=1}^{n} p(x_i, y_i|\theta) \\
&= \prod_{i=1}^{n} p(x_i|y_i, \theta) p(y_i|\theta) \\
&= \prod_{i=1}^{n} w_{y_i} p(x_i|\theta_{y_i}),
\end{aligned}
\tag{4.4}
$$

where $w_j = p_Y(j) = P(Y = j)$. The log-likelihood is

$$
\begin{aligned}
\log p(\mathcal{D}|\theta) &= \sum_{i=1}^{n} \left(\log w_{y_i} + \log p(x_i|\theta_{y_i}) \right) \\
&= \sum_{j=1}^{m} n_j \log w_j + \sum_{i=1}^{n} \log p(x_i|\theta_{y_i}),
\end{aligned}
$$

where n_j is the number of data points in $\mathcal{D}$ generated by the j-th mixing component.

It is useful to observe here that when $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is known, the internal summation operator in Equation (4.3) disappears. More importantly here, it follows that Equation (4.4) can be maximized in a relatively straightforward manner by separating the estimation of $\boldsymbol{w}$ from θ_j's. Let us show how. To find $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$ we can simply find the maximum likelihood estimates for each w_j using the subset of data points coming from distribution j. However, we will make it slightly more complicated and solve a constrained optimization problem, using the method of Lagrange multipliers. Because we know that the

[1] Notice that although the likelihood function $p(\mathcal{D}|\theta)$ has $O(m^n)$ terms, it can be calculated in $O(mn)$ time as a log-likelihood. Thus, calculating the likelihood, when the parameters are known, is tractable.

weights sum to one and that each weight is nonnegative, we must simultaneously deal with equality and inequality constraints. To start we shall first form the Lagrangian function as

$$L(\boldsymbol{w}, \alpha, \boldsymbol{\mu}) = \sum_{j=1}^{m} n_j \log w_j + \alpha \left(\sum_{j=1}^{m} w_j - 1 \right) + \sum_{j=1}^{m} \mu_j w_j$$

where $\alpha \neq 0$ and $\boldsymbol{\mu} \geq \mathbf{0}$ are Lagrange multipliers. Then, by setting

$$\begin{aligned} \frac{\partial}{\partial w_k} L(\boldsymbol{w}, \alpha, \boldsymbol{\mu}) &= 0 \qquad \forall k \in \mathcal{Y} \\ \frac{\partial}{\partial \alpha} L(\boldsymbol{w}, \alpha, \boldsymbol{\mu}) &= 0 \\ w_k \mu_k &= 0 \qquad \forall k \in \mathcal{Y} \quad \text{(the Karush-Kuhn-Tucker conditions)} \end{aligned}$$

we derive that $w_k = -\frac{n_k}{\alpha}$ and $\alpha = -n$.[2] Thus,

$$w_k = \frac{1}{n} \sum_{i=1}^{n} 1(y_i = k),$$

where $1(\cdot)$ is the indicator function defined in Equation (1.6). To find all θ_j's, we cannot proceed without being more concrete about distributions $p(x|\theta_j)$. To do that, we shall assume that each $p(x|\theta_j)$ is an exponential distribution with a parameter λ_j; i.e., $p(x|\theta_j) = \lambda_j e^{-\lambda_j x}$, where $\lambda_j > 0$. We now proceed by setting

$$\frac{\partial}{\partial \lambda_k} \sum_{i=1}^{n} \log p(x_i | \lambda_{y_i}) = 0,$$

for each $k \in \mathcal{Y}$.[3] We obtain that

$$\lambda_k = \frac{n_k}{\sum_{i=1}^{n} 1(y_i = k) \cdot x_i},$$

which is simply the inverse mean over those data points generated by the k-th mixture component. In summary, we observe that if the mixing component designations $\boldsymbol{y}$ are known, the parameter estimation is greatly simplified. This was achieved by decoupling the estimation of mixing proportions and all parameters of the mixing distributions.

In the second hypothetical scenario, suppose that parameters θ are known, and that we would like to estimate the best configuration of the mixture designations $\boldsymbol{y}$ (one may be tempted to call them "class labels"). This task looks like clustering in which cluster memberships need to be determined based on the known set of mixing distributions and mixing probabilities. To do this we can calculate the posterior distribution of $\boldsymbol{Y}$ as

[2] More technically, all $\mu_k = 0$ when $n_k > 0$ and if $n_k = 0$ we can derive that $\mu_k = n$. Since it must hold that $\mu_k \geq 0$, we have a valid solution.

[3] In this case, the Karush-Kuhn-Tucker conditions $\mu_k \lambda_k = 0$ ensure that each $\mu_k = 0$ because $\lambda_k > 0$. By recognizing it, we therefore opted to not write the full Lagrangian and solve it formally. Instead we decided to solve an unconstrained problem and verify in the end that the solutions are in the constraint set.

$$\begin{aligned} p(\boldsymbol{y}|\mathcal{D},\theta) &= \prod_{i=1}^{n} p(y_i|x_i,\theta) \\ &= \prod_{i=1}^{n} \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)} \end{aligned} \tag{4.5}$$

using the Bayes rule. We can now find the best configuration of $\boldsymbol{y}$ out of m^n possibilities. Obviously, because of the i.i.d. assumption each element y_i can be estimated separately and, thus, this estimation can be completed in $O(mn)$ time. The MAP estimate for y_i can be found as

$$\hat{y}_i = \underset{k \in \mathcal{Y}}{\arg\max} \left\{ \frac{w_k p(x_i|\theta_k)}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)} \right\}$$

for each $i \in \{1, 2, \ldots, n\}$.

In reality, neither "class labels" $\boldsymbol{y}$ nor the parameters θ are known. Fortunately, we have just seen that the optimization step is relatively straightforward if one of them is known. Therefore, the intuition behind our algorithm is to form an iterative procedure by *assuming* that either $\boldsymbol{y}$ or θ is known and calculate the other. For example, we can initially pick some value for θ, say $\theta^{(0)}$, and then estimate $\boldsymbol{y}$ by computing $p(\boldsymbol{y}|\mathcal{D},\theta^{(0)})$ as in Equation (4.5). We can refer to this estimate as $\boldsymbol{y}^{(0)}$. Using $\boldsymbol{y}^{(0)}$ we can now refine the estimate of θ to $\theta^{(1)}$ using Equation (4.4). We can then iterate these two steps until convergence. In the case of mixture of exponential distributions, we arrive at the following algorithm:

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Calculate $y_i^{(0)} = \underset{k \in \mathcal{Y}}{\arg\max} \left\{ \frac{w_k^{(0)} p(x_i|\lambda_k^{(0)})}{\sum_{j=1}^{m} w_j^{(0)} p(x_i|\lambda_j^{(0)})} \right\}$ for $\forall i \in \{1, 2, \ldots, n\}$
3. Set $t = 0$
4. Repeat until convergence
   (a) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} 1(y_i^{(t)} = k)$
   (b) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} 1(y_i^{(t)}=k)}{\sum_{i=1}^{n} 1(y_i^{(t)}=k) \cdot x_i}$
   (c) $y_i^{(t+1)} = \underset{k \in \mathcal{Y}}{\arg\max} \left\{ \frac{w_k^{(t+1)} p(x_i|\lambda_k^{(t+1)})}{\sum_{j=1}^{m} w_j^{(t+1)} p(x_i|\lambda_j^{(t+1)})} \right\}$
   (d) $t = t + 1$
5. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

This procedure is close but not quite yet the EM algorithm; rather, it is a version of it referred to as classification EM algorithm (CEM). In the next section we will introduce the EM algorithm.

4.4.2 The expectation-maximization algorithm

The previous procedure was designed to iteratively estimate class memberships and parameters of the distribution. In reality, it is not necessary to compute $\boldsymbol{y}$; after all, we only need to estimate θ. To accomplish this, at each step t, we can compute $p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)})$ to maximize the *expected log-likelihood* of both $\mathcal{D}$ and $\boldsymbol{y}$.

$$\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}] = \sum_{\boldsymbol{y}} \log p(\mathcal{D}, \boldsymbol{y}|\theta) p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}), \tag{4.6}$$

which can be carried out by integrating the log-likelihood function of $\mathcal{D}$ and $\boldsymbol{y}$ over the posterior distribution for $\boldsymbol{y}$ in which the current values of the parameters $\theta^{(t)}$ are assumed to be known. We can now formulate the expression for the parameters in step $t+1$ as

$$\theta^{(t+1)} = \arg\max_{\theta} \left\{ \mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}] \right\}. \tag{4.7}$$

The formula above is all that is necessary to create the update rule for the EM algorithm. Note, however, that inside of it we always have to re-compute $\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$ function because the parameters $\theta^{(t)}$ have been updated from the previous step. We then can perform maximization. Hence the name "expectation-maximization", although it is perfectly valid to think of the EM algorithm as an iterative maximization of expectation from Equation (4.6); i.e., "expectation maximization".

We now proceed as follows

$$\begin{aligned}
\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}] &= \sum_{y_1=1}^{m} \cdots \sum_{y_n=1}^{m} \log p(\mathcal{D}, \boldsymbol{y}|\theta) p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)}) \\
&= \sum_{y_1=1}^{m} \cdots \sum_{y_n=1}^{m} \sum_{i=1}^{n} \log p(x_i, y_i|\theta) \prod_{l=1}^{n} p(y_l|x_l, \theta^{(t)}) \\
&= \sum_{y_1=1}^{m} \cdots \sum_{y_n=1}^{m} \sum_{i=1}^{n} \log \left(w_{y_i} p(x_i|\theta_{y_i}) \right) \prod_{l=1}^{n} p(y_l|x_l, \theta^{(t)}).
\end{aligned}$$

After several simplification steps, that we omit for space reasons, the expectation of the likelihood can be written as

$$\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}] = \sum_{i=1}^{n} \sum_{j=1}^{m} \log \left(w_j p(x_i|\theta_j) \right) p_{Y_i}(j|x_i, \theta^{(t)}),$$

from which we can see that $\boldsymbol{w}$ and $\{\theta_j\}_{j=1}^{m}$ can be separately found. In the final two steps, we will first derive the update rule for the mixing probabilities and then by assuming the mixing distributions are exponential, derive the update rules for their parameters.

To maximize $\mathbb{E}[\log p(\mathcal{D}, \boldsymbol{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$ with respect to $\boldsymbol{w}$, we observe that this is an instance of constrained optimization because it must hold that $\sum_{j=1}^{m} w_j = 1$ and $w_j \geq 0$. We will use the method of Lagrange multipliers; thus, for each $k \in \mathcal{Y}$ we need to solve

$$\frac{\partial}{\partial w_k} \left(\sum_{j=1}^{m} \log w_j \sum_{i=1}^{n} p_{Y_i}(j|x_i, \theta^{(t)}) + \alpha \left(\sum_{j=1}^{m} w_j - 1 \right) \right) = 0,$$

where α is the Lagrange multiplier.[4] It is relatively straightforward to show that

$$w_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{Y_i}(k|x_i, \theta^{(t)}). \tag{4.8}$$

Similarly, to find the solution for the parameters of the mixture distributions, we obtain that

$$\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} p_{Y_i}(k|x_i, \theta^{(t)})}{\sum_{i=1}^{n} x_i p_{Y_i}(k|x_i, \theta^{(t)})} \tag{4.9}$$

for $k \in \mathcal{Y}$. As shown in Equation (4.5), we have

$$p_{Y_i}(k|x_i, \theta^{(t)}) = \frac{w_k^{(t)} p(x_i|\lambda_k^{(t)})}{\sum_{j=1}^{m} w_j^{(t)} p(x_i|\lambda_j^{(t)})}, \tag{4.10}$$

which means that all values of $p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)})$ can be computed and stored as an $n \times m$ matrix. In summary, for the mixture of m exponential distributions, we summarize the EM algorithm by combining Equations (4.8-4.10) as follows:

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Set $t = 0$
3. Repeat until convergence
   (a) $p_{Y_i}(k|x_i, \theta^{(t)}) = \frac{w_k^{(t)} p(x_i|\lambda_k^{(t)})}{\sum_{j=1}^{m} w_j^{(t)} p(x_i|\lambda_j^{(t)})}$ for $\forall(i, k)$
   (b) $w_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{Y_i}(k|x_i, \theta^{(t)})$
   (c) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} p_{Y_i}(k|x_i, \theta^{(t)})}{\sum_{i=1}^{n} x_i p_{Y_i}(k|x_i, \theta^{(t)})}$
   (d) $t = t + 1$
4. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

The convergence can be assessed by computing the likelihood function from Equation (4.3) and terminating the updates once the changes are small enough. We also note that similar update rules can be obtained for different probability distributions; however, separate derivatives have to be found. In summary, in each step t, the EM algorithm performs the following steps:

1. E-step: Compute $p(\boldsymbol{y}|\mathcal{D}, \theta^{(t)})$
2. M-step: Compute $\theta^{(t+1)}$

[4] We here ignore the inequality constraints as we saw in the previous section that the $\boldsymbol{\mu}$ parameters do not affect the outcome of optimization.

Notice the difference between the CEM and the EM algorithms. Although the algorithms are similar in nature, the CEM algorithm "predicts" the class memberships and from them infers the model parameters. On the other hand, the EM algorithm calculates the probabilities of class memberships and uses these probabilities to directly infer the new parameters of the model. By not having to make decisions on the class memberships, the EM algorithm walks differently through the parameter space and generally reaches better decisions, albeit slower than the CEM algorithm.

4.4.3 Identifiability

When estimating the parameters of a mixture, it is possible that for some parametric families one obtains multiple solutions. In other words,

$$\begin{aligned} p(x|\theta) &= \sum_{j=1}^{m} w_j p(x|\theta_j) \\ &= \sum_{j=1}^{m'} w'_j p(x|\theta'_j) \\ &= p(x|\theta') \end{aligned}$$

The parameters are identifiable if

$$\sum_{j=1}^{m} w_j p(x|\theta_j) = \sum_{j=1}^{m'} w'_j p(x|\theta'_j),$$

implies that $m = m'$ for each $j \in \{1, 2, \ldots, m\}$ there exists some $l \in \{1, 2, \ldots, m\}$ such that $w_j = w'_l$ and $\theta_j = \theta'_l$.