Lecture 0a: Tractablility in univariate Bayesian inference

Bayesian inference is a complete and self-consistent methodology that a rational agent can use to extract information from their environment and integrate it with their pre-existing beliefs. The whole of Bayesian statistics can be summarized into a single formula. The formula gives you the only rational method to update belief on the world after you experience a sensation:

$$p(\text{world} \mid \text{measurements}) = \frac{p(\text{measurements} \mid \text{world}) p(\text{world})}{p(\text{measurements})} \tag{1}$$

where $p(\text{world})$ quantifies the agent prior belief on the state of the external world and $p(\text{measurements} \mid \text{world})$ is a model of the sensory experience you would experience if the world was in a given state. The denominator of Bayes theorem is called *model evidence* and it involves a sum over all world states allowed by your model:

$$p(\text{measurements}) = \sum_{\text{All possible states of the world}} p(\text{measurements} \mid \text{world}) p(\text{world}) \, , \tag{2}$$

where the sum can be finite or infinite or be an integral where the world states can be placed in a continuum.

0.0.1 Tractable and intractable Bayesian models

Loosely speaking, a Bayesian inference problem is said to be tractable when the posterior distribution can be computed exactly or approximate with arbitrarily high accuracy using a realistic amount of time and computational resources. This hand-wavy definition can be formalized mathematically using the tools of complexity theory. However, you do not need any advance mathematics to understand that the sum in Eq. 2 is not tractable if there are more world states than atoms in the universe. This could initially sound like a philosophical issue of no practical value. Why would have a realistic model we use in practice have so many possible states? In practice, our models are extreme simplification of reality and their complexity is definitely smaller than most things in the real world. Unfortunately, the number of possible states of almost all machine learning models you will use in your career have enough state to make the number of atoms in the universe into an irrelevant rounding error. The reason is that a state of the world is not a single "thing" but it is a specific "combination of all things". For example, if the world is defined by an old 8 bits RAM chip, the number of possible states is $2^8 = 256$. If instead of a single chip we take the full RAM of a vintage Commodore 64, we would have 2^{64000} of possible states; a number so ridiculously big that calling it astronomically would be a severe overestimation of astronomy.

In general, the state space of a model with M N-valued variables has N^M elements. Superficially it would therefore seem that Bayesian inference is impossible in all but the most trivial small models. However, this is a overly pessimistic attitude since several non-trivial models, while having an extremely large state space, have also a large amount of additional "structure" that makes tractable inference possible. In this lecture, we will discuss some of these tractable models and learn how to compute their posterior distribution.

1 Exact univariate Bayesian analysis

In this section we will discuss inference in Bayesian models with a single either discrete or continuous latent variable. These models are often tractable since, as we discussed above, the biggest source of intractability comes from the exponential growth of states with the number of variables.

We will devote particular attention to analytically tractable models, where the posterior distribution can be expressed as a finite algebraic combination of elementary functions such as powers, exponentially, trigonometric functions and so on. These solutions are said to be expressed in closed-form. A distribution is usually said to be in closed-form if it is expressed with a formula where there are not integrals or infinite sums "left to be solved". For example,

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{3}$$

is said to be in closed form while

$$p(x) = \frac{e^{-x^2/2}}{\int_{-\infty}^{\infty} e^{-y^2/2} \mathrm{d}y} \tag{4}$$

is not. Unfortunately, the language of simple functions is very limited and most interesting distribution defined by pretty looking integrals cannot be expressed in closed-form. Analytic tractability is a rarity that often reflects a very deep structure in the inference problem. However, many more models with less structure are still tractable as they can be approximated arbitrarily well. Unfortunately, most interesting problems are neither analytically tractable nor tractable...

1.1 Discrete inference

We start our treatment with the simplest possible case: discrete inference problems with a single M-valued variables. Clearly, these models are tractable as far as M is small enough. In a discrete problem, we can express the prior as a vector of probabilities:

$$\boldsymbol{\rho}_{\text{prior}} = (p(x = 1), \ldots, p(x = M)) \ , \tag{5}$$

which of course needs to be non-negative valued and sum to one. Similarly, given the data D we can express the likelihood as another non-negative vector:

$$\boldsymbol{l} = (p(D \mid x = 1), \ldots, p(D \mid x = M)) \ . \tag{6}$$

note that the likelihood does not usually sum to one as it is a probability distribution in D but not in x. The model evidence in this case is simply the sum of the entry-wise products between the two vectors:

$$p(D) = \sum_m p(D \mid x = m) p(x = m) \ , \tag{7}$$

which can also be expressed as a dot product

$$p(D) = \boldsymbol{l} \cdot \boldsymbol{\rho}_{\text{prior}} \ . \tag{8}$$

We can now use Bayes formula to compute the posterior distribution:

$$p(x = k \mid D) = \frac{p(D \mid x = k) p(x = k)}{p(D)} \tag{9}$$

$$= \frac{1}{\boldsymbol{l} \cdot \boldsymbol{\rho}_{\text{prior}}} p(D \mid x = k) p(x = k) \ . \tag{10}$$

The posterior can itself be neatly expressed as a vector of probabilities:

$$\boldsymbol{\rho}_{\text{post}} = (p(x = 1 \mid D), \ldots, p(x = M \mid D)) \ . \tag{11}$$

Using this notation, we arrive at a memorable formulation of Bayesian inference in terms of vector operations:

$$\boldsymbol{\rho}_{\text{post}} = \frac{\boldsymbol{l} \odot \boldsymbol{\rho}_{\text{prior}}}{\boldsymbol{l} \cdot \boldsymbol{\rho}_{\text{prior}}} \ , \tag{12}$$

where $\odot$ denotes the entry-wise product.

Clearly, discrete inference is tractable when M is small enough. However, sometimes it can be tractable even when M is exponentially large by exploiting some structure of the problem. The simplest example is when the intersection of the non-zero entries of $\boldsymbol{l}$ and of $\boldsymbol{\rho}_{\text{prior}}$ is tractably small. Specifically, we define as A the set of all indices m such that $p(D \mid x_m) \neq 0$ and $p(x = m) \neq 0$. In this case, the model evidence is

$$p(D) = \sum_{m \in A} p(D \mid x_m) p(x = m) + \sum_{m \notin A} 0 \ , \tag{13}$$

which is tractable is the set of non-zero indices is small. Note that we can exploit this special structure within a superficially intractable problems only if we know where the small set A. In fact, even if we know that A exists, finding it is in itself an intractable problem unless there is additional special structure to lead us like a . This is a general feature of tractable inference. It is something that only happens in problems with a small relevant state space and with an additional "trail" structure that allows us to find the needle in the exponential haystack staring from an arbitrary s. No that this is an extremely stringent property, it implies that most starting states s implicitly contain enough information to track down the exponentially small set A. As you can imagine, tractability is rare.

1.1.1 Example: Medical testing

We can now give a small but perhaps counter-intuitive example of discrete Bayesian inference. A person is randomly chose to be tested for a rare form of infection. The frequency of infected people in the population is 0.1%. We can convert this into the prior probability:

$$p(\text{"I"}) = 0.001 \tag{14}$$

where "I" stands for "infected". The test is ery reliable, it has a false positive rate of 1% and a false negative rate of 1%. This can be summarized in the following likelihood:

$$\begin{aligned} p(\text{"P"} \mid \text{"I"}) &= 0.99 \ , & (15) \\ p(\text{"P"} \mid \text{"NI"}) &= 0.01 \ . & (16) \end{aligned}$$

Assume now that the patient's test came out as positive. What is the probability that the patiant is actually infected? We can work it out using Bayes theorem. First, we compute the model evidence:

$$\begin{aligned} p(\text{"P"}) &= p(\text{"P"} \mid \text{"I"})p(\text{"I"}) + p(\text{"P"} \mid \text{"NI"})p(\text{"NI"}) & (17) \\ &= 0.99 \times 0.001 + 0.01 \times 0.999 \approx 0.01 \ . & (18) \end{aligned}$$

This is the probability of a test coming out as positive regardless whether the person has the infection or not. It is a small number since the infection is rare and the test is reliable. We can now work out the posterior probability:

$$\begin{aligned} p(\text{"I"} \mid \text{"P"}) &= \frac{p(\text{"P"} \mid \text{"I"})p(\text{"I"})}{p(”P”)} & (19) \\ &\approx \frac{0.99 \times 0.001}{0.01} \approx 0.1 \ . & (20) \end{aligned}$$

The patient as only 10% of being infected even if the test was positive! This is a consequence of the rarity of the infection in the general population. Note however that this result is only applicable in a randomized test which is usually not the case in clinical practice. Usually, a patient is tested for a rare disease only if they have symptoms specific enough to suggest the presence of this rare condition. A doctor will not prescribe that test to someone with no symptoms or even to someone with generic symptoms such as headache and nausea. Therefore, in a clinically relevant case the prior probability of infection should be much higher than the frequency in the general population. The relevant prior is instead the posterior of the infection given the set of symptoms:

$$p(\text{"I"} \mid \text{"Orange skin color with blue pimples"}) \,. \tag{21}$$

This is surely a much higher number than the original prior! Note that the fact that the prior is an old posterior is not a coincidence. Bayesian inference is quintessentially iterative since it is a rule about how to update knowledge. We will discuss this in greater details later on.

1.2 Interlude: Information theory

In order to get a deeper understanding of Bayesian inference, it is useful to introduce few basic notions from information theory. Importantly, some of the concepts and techniques we will introduce here will play a central role in the rest of the course. Information theory deals with how information is encoded in probabilistic random sources and how this information can transmitted through noisy channels.

We define the information content of a event x_m (a "message") as the logarithm of the inverse of its probability:

$$\mathfrak{i}_m = \log \frac{1}{\rho_m} = -\log \rho_m \tag{22}$$

Note that this quantity is always non-negative since the logarithm of a number smaller than one is always negative. This definition captures the fact that more information is conveyed by rare events rather than common ones. The sentence "Good morning Joe" contains much less information than "a UFO just landed on my garden and asked me for a lollipop". Of course, you can now argue that the sentence like "vbwe!?fzdfas asdw g5t" contains even more information since it is much, much rarer. This paradox is solved by realizing that the definition of information is meaningful only when coupled with a meaningful definition of distinguishable events. Our minds naturally place all strings such as "vbwe!?fzdfas asdw g5t", "asfasfe njkhukw" and "sdfawetnmkj asdk" in the category of "random string of characters". These strings are not really distinguishable (in the sense that we do not bother distinguishing them) and should therefore be counted as a single event which is much more likely to be used than any individual random string.

The information content is a characteristic of a single event given a probabilistic source. This can be used to define the entropy, which is a property of the whole source (i.e. of the probability distribution). The entropy of a distribution is defined as the average information content:

$$\mathcal{H}\left[\boldsymbol{\rho}\right] = \mathbb{E}_{\boldsymbol{\rho}}[\mathrm{i}_m] \tag{23}$$

$$= -\sum_m \rho_m \log \rho_m \tag{24}$$

$$= \boldsymbol{\rho} \cdot \log \boldsymbol{\rho} \ , \tag{25}$$

where the indeterminate form $0 \log 0$ is defined to be equal to 0. Remember that this quantity is always non-negative since it is an average of non-negative values. It is now easy to check that a deterministic source has zero entropy. Deterministic distribution is defined by a one-hot vector $\boldsymbol{\delta}_k$ which is equal to one for all indices except for k and zero otherwise. It is then easy to see that

$$\mathcal{H}[\boldsymbol{\delta}_k] = \boldsymbol{\delta}_k \cdot \log \boldsymbol{\delta}_k = 0 \ . \tag{26}$$

This makes a lot of sense, you cannot convey any information using a system with only one possible state!
The opposite scenario is a uniform distribution where any message is equally likely. This distribution maximizes the entropy, which is equal to

$$\mathcal{H}[\boldsymbol{u}] = -\sum_m \frac{1}{M} \log \frac{1}{M} = \log M \tag{27}$$

This magic formula, which is engraved in the tombstone of Ludwig Boltzmann, explains much of how our everyday world works (If you are curious, you should pick up a thermodynamics textbook!). More prosaically, if we use base 2 logarithms the formula tells that you can encode a maximum of $\log_2 M$ bits of information in a system of M discernible states. This makes a lot of sense sense since a N bits memory chip has 2^N possible states!

Exercise (Advanced, for ambitious students) Prove that the uniform distribution maximizes the entropy. Hint: This is a constrained optimization problem since we have to make sure that the optimizing vector is a valid probability distribution (i.e. that it is normalized). You can do this by *Lagrangian function* consisting of the original function you wanted to optimize (i.e. the entropy) plus a new variable λ (a *Lagrange multiplier*) multiplied to the constraint:

$$\mathcal{L}(\boldsymbol{\rho}, \lambda) = \mathcal{H}[\boldsymbol{\rho}] + \lambda \left(1 - \sum_m \rho_m \right) \ .$$

You can now use regular calculus to find the (unconstrained) optimizer of the Lagrangian function with respect to $\boldsymbol{\rho}$ and λ. The general Lagrange multiplier technique transforms a N variable optimization with M constraints into a $N + M$ unconstrained optimization.

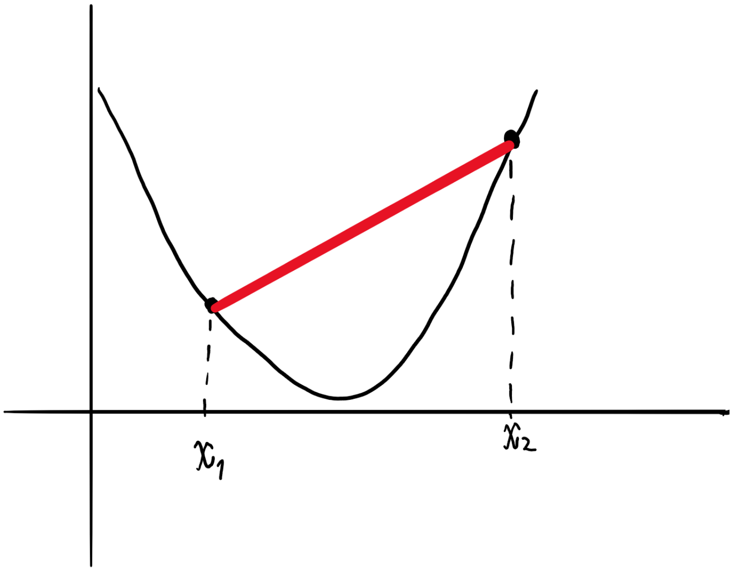

Figure 1: A convex function.

1.2.1 An important tool: Jensen inequality

We can now discuss an important inequality that is of great use in information theory, Bayesian statistics and variational inference. This inequality says that an average of a convex function is always bigger than the convex function of the average. A function f is said to be convex when, for all x_1 and x_2 in the domain of the function and for each $0 \leq a \leq 1$, we have that:

$$f\left(ax_1 + (1-a)x_2\right) \leq af(x_1) + (1-a)f(x_2)\ . \tag{28}$$

Geometrically, this means that the straight line connecting any two points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ is lies always above the function itself. This is sketched in Fig. 1. As you can see, convex functions always smile. The negative of a convex function is said to be concave and it follows the inverse inequality. You can see this by flipping a convex function upside down.

The definition of convexity can be interpreted probabilistically. Since the number a is between zero and one, it can be interpreted as the probability of the event x_1 in a binary setting where only two events x_1 and x_2 are possible. With this interpretation, we can say that a function of a random variable x is convex when the function of the average is always bigger than the average of the function for all possible binary distributions over all possible pair of points in its domain. It is easy to show that this definition implies that the same property is also true for arbitrary non-binary distributions:

$$f\left(\sum_m p_m x_m\right) \leq \sum_m p_m f(x_m) \tag{29}$$

where $\sum_n p_n = 1$. This can be stated more succinctly as follows:

$$f\left(\mathbb{E}[x]\right) \leq \mathbb{E}[f(x)] \tag{30}$$

This is the famous Jensen inequality and it is valid also for expectations over continuous distributions.

For our purposes, we are mostly interested in using this inequality when the function is a logarithm. Logarithms are concave functions, therefore the negative of a logarithm is convex, so that:

$$\log\left(\mathbb{E}[x]\right) \geq \mathbb{E}[\log(x)]\ . \tag{31}$$

Using this result, we can prove that the entropy of the uniform distribution is higher than the entropy of any other distribution

$$\mathcal{H}[\boldsymbol{\rho}] = \sum_k \rho_k \log\frac{1}{\rho_k} \leq \log\left(\sum_k \frac{\rho_k}{\rho_k}\right) = \log M \tag{32}$$

Exercise Use Jensen inequality starting from the definition of convex function using a recursive argument.

1.2.2 Cross-entropy and discrete KL divergence

We can now define one of the most important objects of this course: the Kullback–Leibler (KL) divergence. We begin by defining the cross-entropy between two distributions $\boldsymbol{\rho}$ and $\boldsymbol{q}$ defined over the same event set:

$$\mathcal{H}[\boldsymbol{\rho}, \boldsymbol{q}] = -\boldsymbol{\rho}\cdot\log\boldsymbol{q} \tag{33}$$

$$= -\sum_m \rho_m \log q_m\ . \tag{34}$$

Some of you will probably recognize this formula from the loss functions commonly used in deep learning classification problems. Note that the cross-entropy is well defined only if the support of q (i.e. the set of indices n where q_n is different from zero) is a sub-set of the support of $\boldsymbol{\rho}$.

It is easy to see that the cross entropy $\mathcal{H}[\boldsymbol{\rho}, \boldsymbol{q}]$ is always bigger than the entropy $\mathcal{H}[\boldsymbol{\rho}]$:

$$\mathcal{H}[\boldsymbol{\rho}, \boldsymbol{q}] - \mathcal{H}[\boldsymbol{\rho}] = \sum_m \rho_m \log\frac{\rho_m}{q_m} \tag{35}$$

$$\geq \sum_m \rho_m \left(1 - \frac{q_m}{\rho_m}\right) \tag{36}$$

$$1 - \sum_m q_m = 0\ , \tag{37}$$

where we used the fact that $-\log(x) \geq 1 - x$ for all values of x.
Therefore, the difference between cross-entropy and entropy is always non-negative. Furthermore, it is zero if and only if $\boldsymbol{\rho} = \boldsymbol{q}$. We can therefore define a divergence between distributions using this difference:

$$\begin{aligned} D_{\mathrm{KL}}\left[\boldsymbol{\rho}, \boldsymbol{q}\right] &= \mathcal{H}[\boldsymbol{\rho}, \boldsymbol{q}] - \mathcal{H}[\boldsymbol{\rho}] && (38) \\ &= \sum_m \rho_m \log \frac{\rho_m}{q_m} \ . && (39) \end{aligned}$$

Statistical divergences are a way of measuring the "distance" between probability distributions. Note however that this "distance" is not symmetric since $D_{\mathrm{KL}}\left[\boldsymbol{\rho}, \boldsymbol{q}\right]$ is in general difference from $D_{\mathrm{KL}}\left[\boldsymbol{q}, \boldsymbol{\rho}\right]$. For example, the KL divergence between the uniform distribution $\boldsymbol{u}$ and the deterministic distribution $\boldsymbol{\delta}_k$ is not well defined

$$D_{\mathrm{KL}}\left[\boldsymbol{u}, \boldsymbol{\delta}_k\right] = -\sum_{m=1}^{M} \frac{1}{M} \log\left(M \delta_{k,n}\right) = -\frac{M-1}{M} \log\left(0\right) - \frac{1}{M} \log\left(M\right) = \infty \quad (40)$$

since $\lim_{x \to 0} \log(x) = -\infty$. On the other hand, the reversed divergence is finite:

$$D_{\mathrm{KL}}\left[\boldsymbol{\delta}_k, \boldsymbol{u}\right] = \sum_{m=1}^{M} \delta_{k,n} \log\left(M \delta_{k,n}\right) = \log M \ . \quad (41)$$

The behaviour of the KL divergence can be quite counter-intuitive as it can label as "infinitely distant" some pairs of distributions that we would intuitively think to be very similar. For example, the KL divergence between a distribution $\boldsymbol{\rho}_1$ with all non-zero entries and another distribution $\boldsymbol{\rho}_2$ that is identical to $\boldsymbol{\rho}_1$ except for the probability of the k-th event having probability 0 instead of 0.000001 is infinite!

1.3 The dynamic of sequential inference

In the medical example, we saw that the posterior distribution of a previous inference played the role of prior distribution in the following inference. This is a general feature of Bayesian inference. Assume a situation when at each time point t an agent independently collect a new datum D_t. We denote the likelihood at time t as $\boldsymbol{l}_t = (p(D_t \mid x = 1), \dots, p(D_t \mid x = M))$. We can then define the temporal evolution over the belief of a rational agent using the iterative Bayes rule:

$$\boldsymbol{\rho}_{t+1} = \frac{\boldsymbol{l}_t \odot \boldsymbol{\rho}_{\mathrm{t}}}{\boldsymbol{l}_t \cdot \boldsymbol{\rho}_{\mathrm{t}}} \ . \quad (42)$$

This iterative update can produce several possible stochastic dynamics depending on the likelihood function. However, we can prove several general feature that are true in all iterative inference problems. The most important of these general properties is a sort of reversed law of thermodynamics. Namely, on average the entropy of the rational belief $\boldsymbol{\rho}_t$ is a non-decreasing function of the time index:

$$\mathbb{E}_{D_t}\left[\mathcal{H}[\boldsymbol{\rho}_{t+1}]\right] \leq \mathcal{H}[\boldsymbol{\rho}_t] \ . \quad (43)$$

where
The result can be interpreted as follows: The entropy of the belief reflects the uncertainty of the agent and uncertainty decreases as data is registered. This seems to contradict our previous interpretation of the entropy as the amount of information that can be encoded in the source. Certainly an agent who observed a lot of data has more information than an agent who did not! This is true but it is just a different perspective. The previous interpretation involved using a probabilistic source (in this case the belief of an agent) to encode messages and a very certain agent leaves you little space for this since its belief is sharply concentrated on a few possible events.
Under some conditions, the iterative inference leads the entropy to tend to zero. In this case, the posterior distribution converges to a deterministic distribution:

$$\lim_{t \to \infty} \boldsymbol{\rho}_t = \boldsymbol{\delta}_k \ , \tag{44}$$

where k is the index if the "true state" x_K that generated the data D_t. This deterministic convergence when the likelihood contains information that can potentially discriminate all states x_m. Note that this is usually not the case in real world situations, you can look at a dog as many times as you want but you will never acquire complete information of its atomic and molecular composition since your eyes do not have enough resolution to distinguish the microscopic state.

1.4 General univariate continuous inference

We can now move on to Bayesian inference with a single continuous latent variables. In the continuous case, the model evidence is expressed as an integral:

$$p(D) = \int_a^b p(D \mid x) p(x) \mathrm{d}x \ , \tag{45}$$

where the numbers a and b define the domain of definition (the support) of the distribution. In the Bayesian statistic literature, we are usually sloppy and we leave this domain implicit by writing the integral as

$$p(D) = \int p(D \mid x) p(x) \mathrm{d}x \ . \tag{46}$$

This is not meant to be an indefinite integral! It is just a regular integral where the bounds have been omitted for notational simplicity. We can then express Bayes theorem using the usual formula:

$$p(x \mid D) = \frac{p(D \mid x) p(x)}{p(D)} \ . \tag{47}$$

When the integral in Eq. 46 can be solved exactly, we say that the resulting posterior can be expressed in closed form. For example, consider the inference problem defined by the likelihood over the data point y

$$p(y \mid x) = x e^{-yx}, \quad \text{for } x, y \geq 0 \ , \tag{48}$$

and the prior

$$p(x) = e^{-x}, \quad \text{for } x \geq 0 \ . \tag{49}$$

in this case, the model evidence can be computed using integration by parts:

$$\begin{aligned} p(y) = \int_0^\infty x e^{-xy} e^{-x} \mathrm{d}x &= \int_0^\infty x e^{-x(y+1)} \mathrm{d}x && (50) \\ &= \left[\frac{x e^{-x(y+1)}}{-(y+1)} x \right]_0^\infty - \frac{1}{-(y+1)} \int_0^\infty e^{-x(y+1)} \mathrm{d}x && (51) \\ &= \frac{1}{(y+1)^2} \ . && (52) \end{aligned}$$

We can now write the full posterior distribution as a simple formula:

$$p(x \mid y) = \frac{x e^{-x(y+1)}}{(y+1)^2} \ . \tag{53}$$

This is a stereotypical example of closed-form solution as we expressed the posterior as an algebraic combination of simple functions.

1.4.1 Inference quantization

Only very few inference problem can be solved with a closed-form formula. However, most realistic continuous inference problems are tractable in the since that the posterior can be approximated arbitrarily well with a feasible number of computational steps.

The simplest way to approximate a continuous inference is to quantize (discretize) the range of the variable x. This simply means that we spit the domain of x into M bins (y_n, y_{n+1}) and we assume the values to be indistinguishable within these bins. In practice, this means that we assume the likelihood to be constant within each bin, so that the data cannot provide any information that would allow us to distinguish their values:

$$l_k^{(M)} = p^{(M)}(D \mid x \in (x_k, x_{k+1})) := p(D \mid x_k) \ . \tag{54}$$

This kind of assumption is not unrealistic. All real world measuring devices have a limit to their resolution. The appropriate discrete prior to use in this quantize problem is obtained by integrating the original prior in the bin:

$$\rho_k^{(M)} = \int_{x_k}^{x^{k+1}} p(x) \mathrm{d}x \ . \tag{55}$$

We can now compute the quantized model evidence using Eq. 8:

$$p^{(M)}(D) = \boldsymbol{l}_m \cdot \boldsymbol{\rho}_m^{(q)} = \sum_{m=1}^{M} p(D \mid x_m) \int_{x_m}^{x^{m+1}} p(x) \mathrm{d}x \ . \tag{56}$$

If we take $\Delta x = x_{m+1} - x_m$ to be small compared to how fast the prior changes, we approximate the integrals by assuming that the prior is constant within the bins. In this case, the expression simplifies into

$$p^{(M)}(D) \approx \sum_{m=1}^{M} p(D \mid x_m) p(x_m) \Delta x \ . \tag{57}$$

Note that this approximation becomes better and better as the bin size gets smaller. This finite sum in Eq. 57 converges to the integral $\int_a^b p(D \mid x) p(x) \mathrm{d}x$ is we take the limit of $M \to \infty$ while shrinking the bin size Δx to zero. Therefore, our quantized inference can be used to approximate the original model evidence:

$$\lim_{M \to \infty} p^{(M)}(D) = p(D) \tag{58}$$

1.4.2 Example 1: Temperature measurement

As an example of continuous inference, we consider the case of a body temperature measurement. We decide to measure the temperature using one of those old mercury thermometers where the reading is given by the height of a mercury column. It is common to model the error distribution of measuring devices using a Gaussian whose standard deviation represent the distribution of measurement errors:

$$p(h \mid T) = \mathcal{N}(h; aT, (0.01\text{mm})^2) \tag{59}$$

where the constant $a = 3$ in unit of mm/degrees converts the real temperature value to the mercury column height. We can now assigning a broad prior distribution over the distribution of internal temperatures of a living human.

$$p(T) = \mathcal{N}(T; 37^\circ, (4^\circ)^2) \ . \tag{60}$$

Note that the dispersion is this prior is much wider than the error distribution of the thermometer. Therefore, we do not expect the prior to introduce a meaningful bias the analysis.

Exercise 1 Assume that we got a reading of 13mm. Write a computer program to compute the quantize posterior using bin size of 0.5° and a range from -30 to $+\ 50$.

Exercise 2: Temperature measurement through prankster friend In this second example, we ask a friend to measure your temperature using the same thermometer. However, you know that your friend is a geeky prankster who will tell you the right reading 50% of the times while telling you the number 42 instead in the remaining cases. Write down the quantized likelihood using the same quntization scheme used before. I) Compute the quantized posterior assuming that your friend tells you that the value was 42mm. Compare it with the prior. II) Now compute the quantized posterior if your friend gives you a reading of 13mm. Compare with the posterior obtained in the previous exercise. Discuss the results.

1.4.3 Tractability of univariate continuous inference

When is univariate continuous inference tractable? Clear not in all cases since we can encode any discrete distribution over integers as "step-shaped" continuous distributions and we already saw that discrete distributions can be intractable. A continuous Bayesian inference problem is tractable when the normalization integral

$$\int_a^b p(D \mid x)p(x)\mathrm{d}x \tag{61}$$

can be approximated with arbitrary precision. There are three sources of intractability: I) There can be an unboundedly large region of the x space with non-vanishing values of $p(D \mid x)p(x)$; II) all the non vanishing values are in a tractable region but we do not know its location III) the function $p(D \mid x)p(x)$ changes values very quickly so that we need an exponentially fine binning in order to properly approximate the integral. Fortunately, these sources of intractability are very rare in real world univariate inference problems. In these cases, inference quantization and other numerical solution of the integral should be your "go to" approach.

1.4.4 Interlude: Information theory for continuous variables

Can we generalize the information theory concepts we discussed above to continuous distributions? Yes but it is a bit subtle and the interpretation is is much less straightforward than in the discrete case. Our approach will be to compute the entropy of a quantized distribution and hope that it will converges to a meaningful value when we let the bin size tend to zero. Consider the continuous distribution $p(x)$ with support (a, b)

$$p_k^{(\Delta x)} = \int_{x_k}^{x_k+\Delta x} p(x)\mathrm{d}x \tag{62}$$

The entropy of the quantized distribution is:

$$\begin{aligned}
\mathcal{H}[\boldsymbol{p}^{(\Delta x)}] &= -\sum_k p_k^{(\Delta x)} \log p_k^{(\Delta x)} &(63)\\
&\approx \sum_k p(x_k)\Delta x \log\left(p(x_k)\Delta x\right) &(64)\\
&= \left(-\sum_k p(x_k)\log\left(p(x_k)\right)\Delta x\right) - \left(\sum_k p(x_k)\Delta_k\right)\log\left(\Delta x\right) &(65)
\end{aligned}$$

The first term of this expression converges to the nice-looking integral

$$\mathcal{DH}[p] = -\int_a^b p(x)\log p(x)\mathrm{d}x \tag{66}$$

when $\Delta x \to 0$. This is the so called differential entropy of the distribution. However, for Δx tending to zero, the second term is asymptotic to $-\log \Delta x$

which clearly diverges to infinity. The infinity comes from the fact that you can encode an unbounded number of bits into the infinite decimal expansion of a real number. Fortunately, this "infinite shift" does not dependent on the original distribution. It is therefore meaningful to use the differential entropy as a measure of the information content of a continuous distribution. Note however that the differential entropy can be arbitrarily negative since it is defined as a finite correction to an infinite amount of information. This reflect the fact the value of a probability density can be larger than one.

We can now try to use the same approach to find the continuous limit of the KL divergence:

$$\lim_{\Delta x \to 0} D_{\mathrm{KL}}[\boldsymbol{p}^{(\Delta x)}, \boldsymbol{q}^{(\Delta x)}] = \lim_{\Delta x \to 0} \left(\sum_k p(x_k) \Delta x \log \left(\frac{p(x_k)\Delta x}{q(x_k)\Delta x} \right) \right) \tag{67}$$

$$= \lim_{\Delta x \to 0} \left(\sum_k p(x_k) \log \left(\frac{p(x_k)}{q(x_k)} \right) \Delta x \right) \tag{68}$$

$$= \int_a^b p(x) \log \left(\frac{p(x)}{q(x)} \right) \mathrm{d}x \ . \tag{69}$$

In this case the problematic $\log \Delta x$ terms cancel out and we end up with a well-defined limit without "infinite shifts". The result is that this continuous generalization of the KL divergence has the same properties of the discrete version and it can be interpreted in the very same way.

2 Analytically tractable models, conjugate priors and exponential family

Now that we developed a good general understanding of univariate Bayesian inference in both the discrete and continuous case, we can move on to the analytically tractable examples that are usually discussed at the very beginning of any Bayesian statistics course. These examples are the survivors of an evolutionary selection process that gradually discarded all forms of inference that cannot be quickly computed on pen and paper by a statistician with the help of some pre-computed tables. In fact, Bayesian statistics evolved in an era when computers were not jet available and when statisticians were still mostly busy fighting off saber-toothed tigers.

You could then be of the idea that these analytically tractable models are just oversimplified historical artifacts which should be discarded in the era of cheap computation. However, this attitude is mostly misguided. Evolution, both biological, cultural and historical is not a random process but it is instead directed towards stable attractor points with special properties. In this case, the fundamental property that has been selected by historical evolution is the stability of iterated inference. In the general inference case, the posterior distribution has a very different form than the prior. Since Bayesian statistics is usually used iteratively, posterior distributions are often used as priors for follow-up inference problems. It would be impractical to keep a table of pre-computed evidence integrals for all possible combinations of priors and posteriors. However, this is feasible in the cases where the posterior has the same form of the prior just with different values of the parameters. In these situations, a single table of Bayesian update rules can be used for an arbitrarily long sequence of iterated inference problems. Therefore, Bayesian statisticians were naturally lead toward the concept of conjugate prior distributions. A prior distribution is said to be conjugate to a likelihood when the resulting Bayesian update produces a posterior distribution that has the same form of the prior but with different values of a finite number of parameters. For example, a Gaussian prior can be updated into another Gaussian with different mean and variance. As you can see, in this case Bayes theorem can be turned into a parameter update rule.

2.1 Conjugate priors

We can now formalize the concept of conjugate prior. From this analysis, we will see the reason of the analytical tractability of conjugate models. Consider a distribution of the form:

$$p(x; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \tilde{p}(x; \boldsymbol{\theta}) \tag{70}$$

where $\tilde{p}(x; \boldsymbol{\theta})$ is a non-normalized non-negative function parameterized by a finite vector of parameters $\boldsymbol{\theta}$. The normalization $Z(\boldsymbol{\theta})$ is given by the integral:

$$Z(\boldsymbol{\theta}) = \int \tilde{p}(x; \boldsymbol{\theta}) \mathrm{d}x \tag{71}$$

where I left the boundaries of integration implicit (you have to get use to it!). We said that a distribution of this form is analytically tractable when the normalization integral can be solved for all values of the parameters. A distribution in the form of Eq. 70 is said to be conjugate to a likelihood $p(D \mid x)$ when:

$$p(x \mid D) = \frac{p(D \mid x) p(x; \boldsymbol{\theta})}{p(D)} = \frac{1}{Z(\boldsymbol{\theta}_{\text{new}}(D))} \tilde{p}(x; \boldsymbol{\theta}_{\text{new}}(D)) \ . \tag{72}$$

In other words, if the distribution is conjugate to the likelihood, Bayesian inference can be summarized by the parameter update rule:

$$\mathcal{B}_{\text{update}}[\boldsymbol{\theta}, D] = \boldsymbol{\theta}_{\text{new}}(D) \ . \tag{73}$$

Note that our definition of analytic tractability automatically implies that the posterior can be expressed in closed-form. Therefore, analytic tractability is preserved by Bayesian updates in conjugate models.

2.2 Conjugate priors and data compression

The update rule in conjugate inference problem has a neat interpretation in terms of data compression. In a general iterated Bayesian inference, the expression of the posterior becomes more and more complex as data is observed. Specifically, in general the t-th iterated posterior depends on all the previous t observations. On the other hand, in conjugate models all the information relevant to the inference problem up to the t-th iteration can be summarized by the parameters $\boldsymbol{\theta}_t$. This allows us to compress all the relevant information contained in a potentially enormous dataset into a few numbers without any loss.

2.3 Beta-Binomial inference

We can now discuss the most common examples of conjugate inference problems. This will hopefully make more clear how conjugate inference works in practice. We start from a simple problem where we would like to infer the bias of a coin. We denote the probability of a coin landing on "head" as x. The probability of getting k "heads" on a total of n coin tosses is given by the binomial probability distribution:

$$p(k \mid x, n) = \binom{n}{k} x^k (1-x)^{n-k} \ . \tag{74}$$

We now need to assign a prior on the probability parameter x. The most common choice of prior is given by the beta distribution:

$$p(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \tag{75}$$

where $B(\alpha, \beta)$ is the beta function:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} \mathrm{d}x \tag{76}$$

Is this a closed-form expression? Most people would say so even though we just hid the integral under a name and a symbol! More precisely, we acknowledge that $B(a, b)$ is a useful function to add to our toolbox and we computed tables and developed efficient techniques to approximate it with arbitrary precision without worrying too much about it. If you think that this is cheating, remember that familiar functions such as $\sin(x)$ or e^x are only defined in terms of infinite series and your calculator can only approximate them using tables and efficient techniques. There is nothing objective about closed-form expressions, they are just combination of symbols that we consider simple and recurring enough to not worry about "solving them" further.

We can now attempt to compute the posterior. Our general strategy is to find an expression of the posterior up to any multiplicative term that does not depend on x. The hope is that we will end up with a single expression which we will be able to normalizing by solving the normalization integral. The first step is to evaluate the joint probability:

$$p(x,k) = p(k \mid x)p(x) = \binom{n}{k} x^k (1-x)^{n-k} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \ . \tag{77}$$

This is proportional to the posterior since the model evidence does not depend on x. We can now simplify the expression by dropping all the constants of proportionality:

$$\begin{aligned} p(x \mid k) \propto p(x,k) &\propto x^k (1-x)^{n-k} x^{\alpha-1}(1-x)^{\beta-1} \ . & (78) \\ &= x^{k+\alpha-1}(1-x)^{n-k+\beta-1} \ , & (79) \\ &= x^{\alpha_{\text{new}}-1}(1-x)^{\beta_{\text{new}}-1} \ . & (80) \end{aligned}$$

where $\alpha_{\text{new}} = \alpha + k$ and $\beta_{\text{new}} = \beta + (n-k)$. From the last line, it is clear that our posterior is proportional to a non-normalized beta distribution. This is great as we already know how to normalize this distribution without solving any integral: we just divide by the beta function with the updated parameters. Therefore, we can conclude that:

$$p(x \mid k) = \frac{1}{B(\alpha_{\text{new}}, \beta_{\text{new}})} x^{\alpha_{\text{new}}-1}(1-x)^{\beta_{\text{new}}-1} \ . \tag{81}$$

To summarize, we showed the the beta prior is conjugate to the binomial likelihood and that the parameter update simply consists in summing the number of "heads" to the α parameter and the number of tails to the β parameter. This suggest that α and β should be interpreted as counters that accumulate the number of previously observed "heads" and "tails". This makes sense in the context of data compression. If our aim is to estimate the bias of a coin, only the total number of "heads" and "tails" should matter while the order order of the sequence has no relevance and can therefore be discarded.

2.4 Gaussian inference of the mean

Gaussian distributions are commonly used for modeling the error distribution of measuring devices. In a Gaussian inference problem, we start from a Gaussian likelihood with fixed standard deviation (measurement error) whose mean is given by the mean of a latent variable we with to estimate. For example, the likelihood can give the reading of a thermometer given the real body temperature. If we perform N independent measurements, we can write the likelihood density as follows

$$\begin{aligned} p(D \mid x; \sigma^2) &= \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_n - x)^2/2\sigma^2} & (82) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_n (y_n - x)^2} \ . & (83) \end{aligned}$$

In this case we assumed that the reading in the remoter reading y was calibrated to be a direct estimate of the latent variable x without the kind of conversion factor we used in Example 1.4.2. We can then assign a normal prior distribution to the latent variable x:

$$p(x;\mu_0,\nu_0) = \frac{1}{\sqrt{2\pi\nu_0^2}} e^{-(x-\mu_0)^2/2\nu_0^2} \ . \tag{84}$$

This distribution capture our prior knowledge about the system of interest. In the case of body temperature, it could have mean 37° and standard deviation 4°. The wider is the standard deviation and the lower is the bias we are introducing in the analysis. However, biases can be a very good think as they allow as to reject outlier values of noisy or faulty measuring devices.

We can now compute the exact posterior distribution. We use the same approach that we used in the Beta-binomial model: we evaluate the poster up all constant of proportionality that do not depend on x and then we figure out how to normalize the resulting expression. The first step is always to write down the joint distribution since since it is proportional to the posterior with the constant of proportionality given by the inverse of the model evidence:

$$\begin{aligned}
p(x \mid D) &= \frac{p(D \mid x)p(x)}{p(D)} \propto p(D \mid x)p(x) && (85)\\
&= \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_n (y_n - x)^2} \frac{1}{\sqrt{2\pi\nu_0^2}} e^{-(x-\mu_0)^2/2\nu_0^2} && (86)\\
&\propto e^{-\frac{1}{2\sigma^2}\sum_n (y_n-x)^2 - (x-\mu_0)^2/2\nu_0^2} \ . && (87)
\end{aligned}$$

We can further simplify this expression by expanding the squares in the exponent and noticing that additive constants in the exponent that do not depend on x give rise to irrelevant proportionality factors in the expression:

$$\begin{aligned}
p(x \mid D) &\propto e^{-\frac{x^2 N}{2\sigma^2} + \frac{x}{\sigma^2}\sum_n y_n - \frac{x^2}{2\nu_0^2} + \frac{x\mu_0}{\nu_0^2}} && (88)\\
&= e^{-\frac{x^2}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\nu_0^2}\right) + x\left(\frac{\mu_0}{\nu_0^2} + \frac{1}{\sigma^2}\sum_n y_n\right)} \ . && (89)
\end{aligned}$$

The expression cannot be further simplified. We now have a density of the form:

$$p(x \mid D) \propto e^{-\frac{\eta_2}{2}x^2 + \eta_1 x} \ . \tag{90}$$

This is the so called *natural parameterization* of a Gaussian distribution where η_1 and η_2 are the natural parameters. Bayesian inference is much easier in this parameterization for reasons that will be clear if you read the section about exponential family distributions. The usual mean and standard deviation parameters can be obtained from the natural parameter by completing the squares in the numerator. We then end up with the distribution:

$$p(x \mid D) = \frac{1}{\sqrt{2\pi\nu_{\text{new}}^2}} e^{-(x-\mu_{\text{new}})^2/2\nu_{\text{new}}^2} \tag{91}$$

where

$$\mu_{\text{new}} = \eta_2^{-1}\eta_1 \tag{92}$$

$$= \left(\frac{1}{\nu_0^2} + \frac{N}{\sigma^2}\right)^{-1} \left(\frac{\mu_0}{\nu_0^2} + \frac{\sum_n y_n}{\sigma^2}\right) \tag{93}$$

$$= \left(\frac{\nu_0^{-2}}{\nu_0^{-2} + \sigma^{-2}/N}\right)\mu_0 + \left(\frac{\sigma^{-2}/N}{\nu_0^{-2} + \sigma^{-2}/N}\right)\frac{1}{N}\sum_n y_n \tag{94}$$

and

$$\nu_{\text{new}}^2 = \eta_2^{-1} = \left(\frac{1}{\nu_0^2} + \frac{N}{\sigma^2}\right)^{-1} . \tag{95}$$

These are the Bayesian update rules for a Gaussian distribution. The posterior mean is a trade-off between the prior mean μ_0 and the usual empirical mean $\frac{1}{N}\sum_n y_n$:

$$\mu_{\text{new}} = \lambda\mu_0 + (1-\lambda)\frac{1}{\sum_n} y_n \tag{96}$$

where the trade-off factor

$$\lambda = \left(\frac{\nu_0^{-2}}{\nu_0^{-2} + \sigma^{-2}/N}\right) \tag{97}$$

is always between zero and one and expresses the confidence in the prior knowledge relative to the noise in the measurement. We use this sort of trade-off all the time in real life. If a thermometer gives a body temperature reading of 50° when we are healthy, we will likely assume that the thermometer is broken. In this simple inference case, the trade-off factor does not depend on the specific reading but only in the known noise level of the device. This is a consequence of the fact that we assumed to have absolute knowledge about the distribution of errors. Contrarily to real measurement devices, abstract likelihoods cannot get broken. The inference can be made more realistic by explicitly modeling the possibility of a broken thermometer in the likelihood. For example, we could assume that the device returns a completely random value with probability 0.01. In this case, the inference will automatically filter out unreasonable values without impacting the posterior. Note that the concept of unreasonability would depend on your prior, another example of the kind of things you get in return by introducing a small bias. Unfortunately, the posterior with these more complex likelihoods cannot be obtained in terms of simple formulas. However, that's not too bad, solving difficult inference problems is the main purpose of this course.

2.5 Gaussian inference with linear observation models

In our first thermometer example, we had a conversion factor g in the likelihood which remaps the external temperature to the height of the mercury column. This is an example of linear measuring device, where the degrees of freedom the the device (e.g. the height of the mercury column) has a linear dependency with the underlying quantity we want to measure. Of course, real linear dependencies are always restricted to a reasonable measuring range, real mercury columns do not become kilometers tall if the thermometer is thrown into the sun. We can idealize a linear measuring devices using a Gaussian likelihood with linear observation model $f(x) = ax + b$:

$$p(D \mid x; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_n (y_n - f(x))^2} \ , \tag{98}$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_n (y_n - (ax+b))^2} \ . \tag{99}$$

$$\tag{100}$$

Which can be coupled to the same Gaussian prior given in Eq. 102. The update rule can be computed using the very same approach we used in the regular Gaussian inference case. However, that is not really required as we already solved the problem without knowing it! We can realize that by defining the new variable

$$z = ax + b \ . \tag{101}$$

This is just a linear transformation of our original latent variable, we can therefore write its prior as another Gaussian with shifted mean and variance:

$$p(x; \mu_0, \nu_0) = \mathcal{N}\left(z; \mu_z = a\mu_0 + b, \nu_z = a^2\nu_0\right) \ . \tag{102}$$

We now reduced the problem to a problem of the previous form which we already solved! We can therefore just copy the old solution. For example, the posterior mean over z is

$$\mu_{z\text{new}} = \left(\frac{\nu_z^{-2}}{\nu_z^{-2} + \sigma^{-2}/N}\right)\mu_z + \left(\frac{\sigma^{-2}/N}{\nu_z^{-2} + \sigma^{-2}/N}\right)\frac{1}{N}\sum_n y_n \tag{103}$$

$$= \left(\frac{\nu_0^{-2}a^{-2}}{\nu_0^{-2}a^{-2} + \sigma^{-2}/N}\right)(a\mu_0 + b) + \left(\frac{\sigma^{-2}/N}{\nu_0^{-2}a^{-2} + \sigma^{-2}/N}\right)\frac{1}{N}\sum_n y_n \tag{104}$$

We now just need to re-transform back the mean to the original variable:

$$\mu_{\text{new}} = \left(\frac{\nu_0^{-2}}{\nu_0^{-2} + \frac{\sigma^{-2}}{a^{-2}N}}\right)\mu_0 + \left(\frac{\frac{\sigma^{-2}}{a^{-2}N}}{\nu_0^{-2} + \frac{\sigma^{-2}}{a^{-2}N}}\right)\left(a^{-1}\frac{1}{N}\sum_n y_n - b\right) \ . \tag{105}$$

Done! We solved the more general problem without much extra work! This kind of "lazy" approach is ubiquitous in mathematics, statistics and physics. Most problems in the areas are solved by mapping them to problems we already solved and them transforming them back. A useful skill to learn!

Besides of the happiness of having achieved something with little work, the final result is not disappointing itself as we can easily interpret the formula. The measurement standard deviation σ is is re-scaled by the gain factor. Small errors are magnified if the reading has steep dependency on the underlying variable. Furthermore, measured average needs to be re-scaled down and the bias needs to be subtracted.

2.6 Gaussian inference with non-linear measuring devices

Linear measuring devises are just a convenient simplification. In some situation, measurements have an unavoidable non-linear component in the range of interest. In that case, we can express the likelihood with an arbitrary non-linear function $f(x)$:

$$p(D \mid x; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_n (y_n - f(x))^2} \; . \tag{106}$$

$$\tag{107}$$

Unfortunately, the non-linearity breaks the conjugacy of the inference and therefore the analytical tractability. In fact, if we again define the new variable:

$$z = f(z) \; , \tag{108}$$

we find out that the resulting transformed distribution is not Gaussian since the shape is distorted by the non linear function:

$$p(z) = \left| \frac{\mathrm{d}f(x)}{\mathrm{d}x} \right|^{-1} \mathcal{N}\left(f^{-1}(z); \mu_0, \nu_0\right) \; , \tag{109}$$

where we assume the function to be invertible so that there is a one-to-one mapping between the latent variable and the device reading. The resulting distribution is not Gaussian and it is therefore not conjugate to the Gaussian likelihood. However, if the function f does not vary too wildly the non-linear inference is still tractable as we can approximate the posterior with arbitrary precision using quantization. Unfortunately, in this case we do not get the deep insights we usually obtain by analyzing the parameters of the posterior. The real world can be an ugly place at times.

3 Advanced: The exponential family and the general theory of Bayesian conjugacy

In this section, we will gain deeper understanding of analytically tractable inference and conjugate priors by introducing the general theory of exponential family inference. This material is more abstract than the rest of the chapter and it can be skipped at a first read.

3.1 The exponential family

Do all likelihood functions have a conjugate prior? The answer is no unfortunately, conjugate priors can only be found for likelihood that are in the exponential family. However, the theory of exponential families gives us an automatic algorithm to find the conjugate prior and all posterior calculations are much simpler when we use their natural parameterization. An exponential family likelihood has the following form:

$$p(y \mid \boldsymbol{\eta}) = e^{-\boldsymbol{\eta}^T \boldsymbol{\Phi}(y) + F(\boldsymbol{\eta})} \ , \tag{110}$$

where $\boldsymbol{\eta}$ is a vector of so called *natural parameters* and $\boldsymbol{\Phi}(y)$ is a vector *sufficient statistics*. $F(\boldsymbol{\eta})$ is usually referred to as the free energy and is defined as

$$F(\boldsymbol{\eta}) = -\log \int e^{-\boldsymbol{\eta}^T \boldsymbol{\Phi}(y)} \mathrm{d}y \ .$$

.

Exponential family likelihoods are very special as they can compress the information contained in an arbitrary large dataset concerning the parameters $\boldsymbol{\eta}$ into a fixed number of quantities without information loss. In fact, the joint probability of a dataset $\{y_n\}$ of independently sampled datapoints is

$$\begin{aligned} p(D \mid \boldsymbol{\eta}) &= \prod_n^N \exp\left(-\boldsymbol{\eta}^T \boldsymbol{\Phi}(y_n) + F(\boldsymbol{\eta})\right) && (111) \\ &= \exp\left(-\sum_n^N \boldsymbol{\eta}^T \boldsymbol{\Phi}(y_n) + NF(\boldsymbol{\eta})\right) && (112) \\ &= \exp\left(-N\boldsymbol{\eta}^T \left(\frac{1}{N}\sum_n^N \boldsymbol{\Phi}(y_n)\right) + NF(\boldsymbol{\eta})\right) \ . && (113) \end{aligned}$$

You can see that the joint distribution depends on the dataset only through the average of the sufficient statistics (hence the name). This means that the any dataset can be summarized with a fixed finite number of values . This is a fundamental property that allows for closed-form Bayesian inference.

Exercises 1 The expectation parameters of an exponential family distribution are defined as follows:

$$m = \mathbb{E}[\boldsymbol{\Phi}(y)] = \int \boldsymbol{\Phi}(y) e^{-\boldsymbol{\eta}^T \boldsymbol{\Phi}(y) + F(\boldsymbol{\eta})} \mathrm{d}x \tag{114}$$

I) Show that the expectation parameters are the derivative of the free energy with respect to the natural parameters:

$$m = \frac{\partial F(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \ . \tag{115}$$

Example 1 The exponential distribution

$$p(x \mid \lambda) = e^{-\lambda x + \log \lambda} \ . \tag{116}$$

is an exponential family distribution with a single natural parameter λ, with sufficient statistic $\boldsymbol{\Phi}(y) = x$ and $F(\lambda) = \log \lambda$. The expectation parameter of the distribution is

$$m = \frac{\partial \log \lambda}{\partial \lambda} = 1/\lambda$$

Example 2 The normal distribution

$$p(x \mid \lambda) = e^{-\eta_1 x - \eta_2 x^2 + F(\eta_1, \eta_2)} \ . \tag{117}$$

is a two parameters exponential family distribution with x and x^2 as sufficient statistics. The free energy is

$$F(\eta_1, \eta_2) = -\frac{\eta_1^2}{4\eta_2} + 1/2 \log \eta_2 - 1/2 \log \pi \ .$$

The more familiar mean parameter can be obtained from the natural parameters as follows

$$\mu = \frac{\partial F(\eta_1, \eta_2)}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} \ .$$

Similarly, the variance of the distribution is

$$\sigma^2 = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2 = \frac{\partial F(\eta_1, \eta_2)}{\partial \eta_2} - \mu^2 = \frac{1}{2\eta_2}$$

3.2 Conjugate priors for exponential family models

Any exponential family likelihood has a conjugate prior over the natural parameters of the following form:

$$p(\boldsymbol{\eta} \mid \boldsymbol{\gamma}, N_0) = \frac{1}{Z(\boldsymbol{\gamma}, N_0)} e^{-\boldsymbol{\eta}^T \boldsymbol{\gamma} + N_0 F(\boldsymbol{\eta})} \ , \tag{118}$$

where γ_0 and n_0 are hyper-parameters and $Z(\gamma_0, t_t, n_0)$ is a normalization factor:

$$Z(\gamma_0, N_0) = \int e^{-\boldsymbol{\eta}^T \boldsymbol{\gamma} + N_0 F(\boldsymbol{\eta})} \mathrm{d}\boldsymbol{\eta} \tag{119}$$

We assume to have a formula for $Z(\boldsymbol{\gamma}, t_t, n_0)$ valid for all possible values of the hyper parameters.

As usual, we can now use Bayes theorem with Eq. 3.1 as likelihood and Eq. 3.2 as its conjugate prior. We will ignore all terms that do not depend on $\boldsymbol{\eta}$ as we can re-normalize the distribution at the end:

$$\begin{aligned} p(\boldsymbol{\eta} \mid D) &\propto \prod_n^N p(y_n \mid \boldsymbol{\eta}) p(\boldsymbol{\eta} \mid \boldsymbol{\gamma}, N_0) & (120) \\ &\propto \exp\left(-\boldsymbol{\eta}^T \sum_n^N \boldsymbol{\Phi}(y_n) + N F(\boldsymbol{\eta}) - \boldsymbol{\eta}^T \boldsymbol{\gamma} + N_0 F(\boldsymbol{\eta})\right) & (121) \\ &= \exp\left(-\boldsymbol{\eta}^T \left(\sum_n^N \boldsymbol{\Phi}(y_n) + \boldsymbol{\gamma}\right) + (N + N_0) F(\boldsymbol{\eta})\right) \ . & (122) \end{aligned}$$

You can see that the resulting distribution has the same form of the prior with updated parameters and can be normalized using our formula for $Z(\gamma, N)$:

$$p(\boldsymbol{\eta} \mid D) = \frac{1}{Z(\boldsymbol{\gamma}_+, N_+)} e^{-\boldsymbol{\eta}^T \boldsymbol{\gamma}_+ + N_+ F(\boldsymbol{\eta})} \tag{123}$$

where $\boldsymbol{\gamma}_+ = \boldsymbol{\gamma} + \sum_n^N \boldsymbol{\Phi}(y_n)$ and $N_+ = N_0 + N$.
TODO: Predictive distribution

Example 1 The conjugate prior of the rate parameter λ of an exponential distribution is a gamma distribution:

$$p(\lambda \mid \gamma_0, N_0) = \frac{\gamma^{N_0+1}}{N_0!} e^{-\lambda\gamma + N_0 \log \lambda} \tag{124}$$

Given a dataset D, the posterior is another gamma distributions with updated parameters:

$$p(\lambda \mid D) = \frac{\gamma_+^{N_1+1}}{N_1!} e^{-\lambda\gamma_+ + N_+ \log \lambda} \tag{125}$$

where $N_+ = N_0 + N$ and $\gamma_+ = \gamma + \sum_n^N y_n$. The parameter N_0 can be interpreted as the number of "prior observations". This makes particular sense in an iterated setting where the old posterior is used as the new prior. The posterior distribution of the expectation parameter m can be obtained using the change of variable formula and is an inverse gamma distribution:

$$p(m \mid D) = \frac{\gamma_+^{N_+ +1}}{N_+!} e^{-\gamma_1/m - (N_+ - 2)\log m} \tag{126}$$

The corresponding update for the mean of the expectation parameter m is

$$\mu_+ = \frac{\gamma_+}{N_+} = \frac{N_0}{N_0 + N}\mu + \frac{N}{N_0 + N}\left(\frac{1}{N}\sum_n^N y_n\right) .$$

As you can see, the posterior mean is an average of the prior mean μ_0 and the dataset empirical mean. This is a general property of posterior updates for expectation parameters.

Example 2: Gaussian inference revisited The conjugate prior of the natural parameters η_1 of a normal distribution with known η_2 is

$$p(\eta_1 \mid \gamma_1, N_0, \eta_2) = \frac{1}{Z_N(m, N_0)} e^{-\eta_1\gamma_1 - \frac{N_0}{4\eta_2}\eta_1^2} . \tag{127}$$

This is another normal distribution with normalization:

$$Z_N(\gamma_1, N_0) = 2\sqrt{\frac{\eta_2\pi}{N_0}} e^{\eta_2\gamma_1^2/N_0} . \tag{128}$$

Given a dataset D, the posterior is another normal distribution with updated parameters:

$$\gamma_{1+} = \gamma_1 + \sum_{n}^{N} y_n \ , \tag{129}$$

$$N_+ = N_0 + N \ . \tag{130}$$

Using the change of variables formula, we can also derive the update rule for the mean of the expectation parameter $\mu = -\eta_1/2\eta_2$:

$$m_+ = \frac{N_0}{N_0 + N} m + \frac{N}{N_0 + N} \left(\frac{1}{N} \sum_{n}^{N} y_n \right) \ ,$$

where N_0 should again be interpreted as the number of prior observations.

Exercise 2 Compute the Bayesian update rule of the posterior parameters η_2 with known η_1.