1 A Probability Primer

Kenji Doya and Shin Ishii

1.1 What Is Probability?

The subtitle of this book is "Probabilistic Approaches to Neural Coding," so, to start with, we have to be clear about what is probability [1].

A classical notion of probability is the so-called frequentist view. If you toss a coin or roll a die infinitely many times, the ratio of having a particular outcome among all possible outcomes would converge to a certain number between zero and one, and that is the probability. An alternative idea of probability is the "Bayesian" view [2], which regards probability as a measure of belief about the predicted outcome of an event.

There has been a long debate between the two camps; the frequentists refuse to include a subjective notion like "belief" into mathematical theory. Bayesians say it is OK, as long as the way a belief should be updated is given objectively [3]. The Bayesian notion of probability fits well with applied scientists' and engineers' needs of mathematical underpinnings for measurements and decisions. As we will see in this book, the Bayesian notion turns out to be quite useful also in understanding how the brain processes sensory inputs and takes actions.

Despite the differences in the interpretation of probability, most of the mathematical derivation goes without any disputes. For example, the Bayes theorem, at the core of the Bayesian theory of inference, is just a straightforward fact derived from the relationship between joint probability and conditional probability, as you will see below.

1.1.1 Probability Distribution and Density

We consider a random variable X, which can take either one of discrete values $x_1, ..., x_N$ or continuous values, for example, $x \in R^n$. We denote by $P(X = x)$, or just $P(x)$ for short, the probability of the random variable X taking a particular value x.

For discrete random variables, $P(X)$ is called *the probability distribution* function. The basic constraint for probability distribution function is non-negativity and unity, i.e.,

$$P(x_i) \geq 0, \qquad \sum_{i=1}^{N} P(x_i) = 1. \tag{1.1}$$

If X takes a continuous value, its probability of taking a particular value is usually zero, so we should consider a probability of X falling in a finite interval $P(X \in [x1, x2])$. Here $P(X)$ gives a *probability density* function, whose constraint is given by

$$P(x) \geq 0, \qquad \int_X P(x)dx = 1. \tag{1.2}$$

Here the integral is taken over the whole range of the random variable X.

Despite these differences, we often use the same notation $P(X)$ for both probability distribution and density functions, and call them just *probability* for convenience. This is because many of the mathematical formulas and derivations are valid for both discrete and continuous cases.

1.1.2 Expectation and Statistics

There are a number of useful quantities, called *statistics*, that characterize a random variable. The most basic operation is to take an *expectation* of a function $f(X)$ of a random variable X following a distribution $P(X)$

$$E_{P(X)}[f(X)] = \sum_{i=1}^{N} P(x_i)f(x_i), \tag{1.3}$$

or a density $P(X)$ as

$$E_{P(X)}[f(X)] = \int_X P(x)f(x)dx. \tag{1.4}$$

We often use shorthand notations $E_X[\ ]$ or even $E[\ ]$ when the distribution or density that we are considering is apparent. Table 1.1 is a list of the most popular statistics.

1.1.3 Joint and Conditional Probability

If there are two or more random variables, say X and Y, we can consider their *joint probability* of taking a particular pair of values, $P(X = x, Y = y)$. We can also consider a *conditional probability* of X under the condition that Y takes a particular value y, $P(X = x|Y = y)$.

Table 1.1 Most popular statistics

name	notation	definition
mean	$< X >, \mu_X$	$E[X]$
variance	$Var[X], \sigma_X^2$	$E[(X - E[X])^2] = E[X^2] - E[X]^2$
covariance	$Cov[X, Y]$	$E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
correlation	$Cor[X, Y]$	$\frac{Cov[X,Y]}{E[X]E[Y]}$

The joint and conditional probabilities have a natural relationship

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X). \tag{1.5}$$

When we start from the joint probability $P(X, Y)$, $P(X)$ and $P(Y)$ are derived by summing or integrating the two-dimensional function toward the margin of the X or Y axis, i.e.

$$P(X) = \sum_{i=1}^{N} P(X, Y = y_i), \tag{1.6}$$

or

$$P(X) = \int_Y P(X, Y = y)dy, \tag{1.7}$$

so they are often called *marginal probablity*.

1.1.4 Independence and Correlation

When the joint probability is just a product of two probabilities, i.e.,

$$P(X, Y) = P(X)P(Y), \tag{1.8}$$

the variables X and Y are said to be *independent*. In this case we have

$$P(X|Y) = P(X), \qquad P(Y|X) = P(Y).$$

Otherwise we say X and Y are dependent.

A related but different concept is *correlation*. We say two variables are uncorrelated if

$$E[XY] = E[X]E[Y]. \tag{1.9}$$

In this case the covariance and correlation are zero.

If two variables are independent, they are uncorrelated, but the reverse is not true. Why? Let's imagine a uniform probability $P(X, Y)$ over a rhombus around the origin of $X - Y$ space. From symmetry, X and Y are obviously uncorrelated, but the marginal probabilities $P(X)$ and $P(Y)$ are triangular, so their product will make a pyramid rather than a flat rhombus, so X and Y are dependent.

1.2 Bayes Theorem

From the two ways of representing the joint probability (1.5), we can relate the two conditional probabilities by the following equation:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}, \tag{1.10}$$

as long as $P(Y)$ never becomes exactly zero. This simple formula is famous as *the Bayes theorem* [2]. The Bayes theorem is just a way of converting one conditional probability to the other, by reweighting it with the relative probability of the two variables. How can we be so excited about this?

This is quite insightful when we use this theorem for interpretation of sensory data, for example,

$$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypothesis)}{P(data)}.$$

Here, the Bayes theorem dictates how we should update our belief of a certain hypothesis, $P(hypothesis)$ based on how well the acquired data were predicted from the hypothesis, $P(data|hypothesis)$. In this context, the terms in the Bayes theorem (1.10) have conventional names: $P(X)$ is called the *prior probability* and $P(X|Y)$ is called the *posterior probability* of X given Y. $P(Y|X)$ is a *generative model* of observing Y under hypothesis X, but after a particular observation is made it is called the *likelihood* of hypothesis X given data y.

The marginal probability $P(Y)$ serves as a normalizing denominator so that the sum of $P(X|Y)$ for all possible hypotheses becomes unity. It appears as if the marginal distribution is there just for the sake of bookkeeping, but as we will see later, it sometimes give us insightful information about the quality of our inference.

1.3 Measuring Information

Neuroscience is about how the brain processes information. But how can we define "information" in a quantitative manner [4]? Let us consider how informative is an observation of a particular value x for a random variable X with probability $P(X)$. If $P(X = x)$ is high, it is not so surprising, but if $P(X = x)$ is close to zero, it is quite informative. The best way to quantify the *information* or "surprise" of an event $X = x$ is to take the logarithm of the inverse of the probability

$$\log \frac{1}{P(X = x)} = -\log P(X = x). \tag{1.11}$$

Information is zero for a fully predicted outcome x with $P(X = x) = 1$, and increases as $P(X = x)$ becomes smaller. The reason we take the logarithm is

that we can measure the information of two independent events x and y, with joint probability $P(x, y) = P(y)P(y)$, by the sum of each event, i.e.

$$\log \frac{1}{P(x,y)} = \log \frac{1}{P(x)P(y)} = \log \frac{1}{P(x)} + \log \frac{1}{P(y)}.$$

It is often convenient to use a binary logarithm, and in this case the unit of information is called a *bit*.

1.3.1 Entropy

By observing repeatedly, x should follow $P(X)$, so the average information we have from observing this variable is

$$H(X) = E[-\log P(X)] = \sum_X -P(X) \log P(X), \tag{1.12}$$

which is called *the entropy* of X. Entropy is a measure of randomness or uncertainty of the distribution $P(X)$, since the more random the distribution, the more information we gather by observing its value. For instance, entropy takes zero for a deterministic variable (as $H(X) = 0$ for $P(X = x) = 1$ and $P(X \neq x) = 0$), and takes the largest positive value $\log N$ for a uniform distribution over N values.

1.3.2 Mutual Information

In sensory processing, it is important to quantify how much information the sensory input Y has about the world state X. A reasonable way is to ask how much uncertainty about the world X decreases by observing Y, so we take the difference in the entropy of $P(X)$ and $P(X|Y)$,

$$I(X;Y) = H(X) - H(X|Y), \tag{1.13}$$

where $H(X|Y)$ is the *conditional entropy*, given by the entropy of conditional distribution $P(X|Y = y)$ averaged over the probability of observation $P(Y = y)$,

$$H(X|Y) = E_{P(Y)}[E_{P(X|Y)}[-\log P(X|Y)]] = \sum_Y P(Y) \sum_X -P(X|Y) \log P(X|Y). \tag{1.14}$$

$I(X;Y)$ is called the *mutual information* of X and Y. It is symmetric with respect to X and Y. This can be confirmed by checking that the entropy of the joint probability $P(X, Y) = P(Y|X)P(X) = P(X|Y)P(Y)$ is given by

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y), \tag{1.15}$$

and hence the mutual information can be presented in three ways:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y). \tag{1.16}$$

1.3.3 Kullback-Leibler Divergence

We often would like to measure the difference in two probability distributions, and the right way to do it is by information. When we observe an event x, its information depends on what probability distribution we assume for the variable. The difference in information with distributions $P(X)$ and $Q(X)$ is

$$\log \frac{1}{Q(x)} - \log \frac{1}{P(x)} = \log \frac{P(x)}{Q(x)}.$$

If x turns out to follow distribution $P(X)$, then the average difference is

$$D(P;Q) = E_{P(X)} \left[\log \frac{P(x)}{Q(x)} \right] = \sum_X P(x) \log \frac{P(x)}{Q(x)}, \tag{1.17}$$

which is called *the Kullback-Leibler (KL) divergence*. This is a good measure of the difference of two distributions, but we cannot call it "distance" because it does not usually satisfy the symmetry condition, i.e., $D(P,Q) \neq D(Q,P)$.

1.4 Making an Inference

Let us now consider the process of perception in a Bayesian way. The brain observes sensory input Y and makes an estimate of the state of the world X.

1.4.1 Maximum Likelihood Estimate

The mechanics of the sensory apparatus determines the conditional probability $P(Y|X)$. One way of making an inference about the world is to find the state X that maximizes the likelihood $P(Y = y|X)$ of the sensory input y. This is called the *maximum likelihood (ML)* estimate. Although the ML estimate is quite reasonable and convenient, there are two possible drawbacks. First, in the world, there are more probable and less probable states, so inference just by the present sensory input may not be the best thing we can do. Second, using just a single point estimate of X can be dangerous because it neglects many other states that are nearly likely.

1.4.2 Maximum a Posteriori Estimate

This is why the Bayes theorem can be useful in perceptual inference. If we express the probability of different world states as a prior probability $P(X)$, we can combine the sensory information and this prior information according to the Bayes therorem:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}. \tag{1.18}$$

If we put aside the normalizing denominator $P(Y)$, the posterior probability $P(X|Y)$ of the world state X given sensory input Y is proportional to the product of the likelihood $P(Y|X)$ and the prior probability $P(X)$. The state X that maximizes the posterior probability is called *the maximum a posterioir (MAP)* estimate.

1.4.3 Bayesian Estimate

The MAP estimate can incorporate our prior knowledge about the world, but it still is a point estimate. We can instead use the full probability distribution or density of the posterior $P(X|Y)$ as our estimate. For example, if we make a decision or motor action based on the estimated world state X, how sharp or flat is the posterior distribution gives us the confidence of our estimate. When the distribution is wide or even has multiple peaks, we can average the corresponding outputs to make a more conservative decision rather than just using a single point estimate.

1.4.4 Bayes Filtering

A practically important way of using the posterior probability is to use it as the prior probability in the next step. For example, if we make multiple independent sensory observations

$$y = (y_1, y_2, ..., y_t),$$

the likelihood of a state given the sequence of observations is the product

$$P(y|X) = P(y_1|X)P(y_2|X)...P(y_t|X).$$

The posterior is given by

$$P(X|y) = \frac{P(y_1|X)P(y_2|X)...P(y_t|X)P(X)}{P(y)}, \tag{1.19}$$

but this can be recursively computed by

$$P(X|y_1, ..., y_t) \propto P(y_t|X)P(X|y_1, ..., y_{t-1}). \tag{1.20}$$

Here, $P(X|y_1, ..., y_{t-1})$ is the posterior of X given the sensory inputs till time $t-1$ and serves as the prior for further estimation at time t.

So far we assumed that the world state X stays the same, but what occurs if the state changes while we make sequential observations? If we have the knowledge about how the world state would change, for example, by a state transition probability $P(X_t|X_{t-1})$, then we can use the posterior at time $t-1$ multiplied by this transition probability as the new prior at t:

$$P(X_t|y_1, ..., y_{t-1}) \propto P(X_t|X_{t-1})P(X_{t-1}|y_1, ..., y_{t-1}). \tag{1.21}$$

Table 1.2 Popular probability distribution and density functions

name	definition	range	mean	variance
Binomial	$\binom{N}{x}\alpha^x(1-\alpha)^{N-x}$ $\binom{N}{x} = \frac{N!}{(N-x)!x!}$	$x = 0, 1, ..., N$	$N\alpha$	$N\alpha(1-\alpha)$
Poisson	$\frac{1}{x!}\alpha^x e^{-\alpha}$	$x = 0, 1, 2, ...$	α	α
Gaussian or normal	$\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	$x \in R$	μ	σ^2
Gamma	$\frac{1}{\Gamma(a)}b^a x^{a-1}e^{-bx}$ $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$ ($\Gamma(a) = a!$ if a is an integer)	$x \geq 0$	$\frac{a}{b}$	$\frac{a}{b^2}$

Thus the sequence of the Bayesian estimation of the state is given by the following iteration:

$$P(X_t|y_1, ..., y_t) \propto P(y_t|X_t)P(X_t|X_{t-1})P(X_{t-1}|y_1, ..., y_{t-1}). \tag{1.22}$$

This iterative estimation is practically very useful and is in general called *the Bayes filter*. The best known classical example of the Bayes filter is the *Kalman filter*, which assumes linear dynamics and Gaussian noise. More recently, a method called *particle filter* has been commonly used for tasks like visual tracking and mobile robot localization [5].

1.5 Learning from Data

So far we talked about how to use our knowledge about the sensory transformation $P(Y|X)$ or state transition $P(X_t|X_{t-1})$ for estimation of the state from observation. But how can we know these transformation and transition probabilities? The brain should *learn* these probabilistic models from experience.

In estimating a probablistic model, it is convenient to use a *parameterized family* of distributions or densities. In this case, the process of learning, or system identification, is regarded as the process of *parameter estimation*. Table 1.2 is a list of popular parameterized distribution and density functions.

When we make an estimate of the parameter, we can use the same principle as we did in the world state estimation above. For example, when the observation Y is a linear function of the state X with Gaussian noise, we have a parameterized model

$$P(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-wx)^2}{2\sigma^2}},$$

where $\theta = (w, \sigma)$ is a parameter vector. From the set of input-output observations $\{(x_1, y_1), ..., (x_T, y_T)\}$, we can derive a maximum likelihood estimation

by searching for the parameter that maximizes

$$P(y_1, ..., y_T|x_1, ..., x_T, \theta) = \prod_{t=1}^{T} P(y_t|x_t, \theta).$$

A convenient way of doing ML estimation is to maximize the log-likelihood:

$$\log P(y_1, ..., y_T|x_1, ..., x_T, \theta) = \sum_{t=1}^{T} \log P(y_t|x_t, \theta) = \sum_{t=1}^{T} -\frac{(y_t - wx_t)^2}{2\sigma^2} - T \log \sqrt{2\pi}\sigma.$$

From this, we can see that finding the ML estimate of the linear weight w is the same as finding the *least mean-squared error (LMSE)* estimate that minimizes the mean-squared error

$$E = \frac{1}{T} \sum_{t=1}^{T} (y_t - wx_t)^2.$$

1.5.1 Fisher Information

After doing estimation, how can we be certain about an estimated parameter $\hat{\theta}$? If the likelihood $P(Y|\theta)$ is flat with respect to the parameter θ, it would be difficult to make a precise estimate. The *Fisher information* is a measure of the steepness or curvature of the likelihood:

$$I_F(\theta) = E_Y \left[\left(\frac{\partial \log P(Y|\theta)}{\partial \theta} \right)^2 \right] = E_Y \left[-\frac{\partial^2 \log P(Y|\theta)}{\partial \theta^2} \right]. \tag{1.23}$$

A theorem called *Cramér-Rao inequality* gives a limit of how small the variance of an *unbiased* estimate $\hat{\theta}$ can be, namely,

$$Var(\hat{\theta}) \geq I_F(\hat{\theta})^{-1}. \tag{1.24}$$

For example, after some calculation we can see that the Fisher information matrix for a Gaussian distribution with parameters $\theta = (\mu, \sigma^2)$ is

$$I_F(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

If data $Y = (y_1, ..., y_T)$ are given by repeated measures of the same distribution, from $\log P(Y|\theta) = \sum_{t=1}^{T} \log P(y_t|\theta)$, the Fisher information is T times that of a single observation. Thus Cramér-Rao inequality tells us how good estimate of the mean μ we can get from the observed data Y depends on the variance and number of observations, $\frac{\sigma^2}{T}$.

1.5.2 Bayesian Learning

We can of course use not only ML, but MAP or Bayesian estimation for learning parameter θ, for example, for the sensory mapping model $P(Y|X,\theta)$ by

$$P(\theta|X,Y) = \frac{P(Y|X,\theta)P(\theta)}{P(Y|X)}. \tag{1.25}$$

In the above linear example, if we have a prior knowledge that the slope is not so steep, we can assume a Gaussian prior of w,

$$P(w) = \frac{1}{\sqrt{2\pi}\sigma_w} e^{-\frac{||w||^2}{2\sigma_w^2}}.$$

Then the log posterior probability is

$$\log P(w|Y,X) = \sum_{t=1}^{T} -\frac{(y_t - wx_t)^2}{2\sigma^2} - \frac{||w||^2}{2\sigma_w^2} - T\log\sqrt{2\pi}\sigma - \log\sqrt{2\pi}\sigma_w - \log P(Y|X),$$

so maximizing it with respect to w is the same as minimizing the least mean-squared error with a penalty term

$$E = \frac{1}{T}\sum_{t=1}^{T} \frac{(y_t - wx_t)^2}{2\sigma^2} + \frac{||w||^2}{2T\sigma_w^2}.$$

Such estimation with additional regularization terms is used to avoid extreme solutions, often in an adhoc manner, but the Bayesian framework provides a principled way of how to design them [6].

1.5.3 Marginal Likelihood

The normalizing denominator $P(Y|X)$ of the posterior distribution (1.25) is given by integrating the numerator over the entire range of the paremeter

$$P(Y|X) = \int P(Y|X,\theta)P(\theta)d\theta, \tag{1.26}$$

which is often called *marginalization*. This is a hard job in a high-dimensional parameter space, so if we are just interested in finding a MAP estimate, it is neglected.

However, this marginal probability of observation Y given X, or *marginal likelihood*, conveys an important message about the choice of our prior $P(\theta)$. If the prior distribution is narrowly peaked, it would have little overlap with the likelihood $P(Y|X,\theta)$, so the expectation of the product will be small. On the other hand, if the prior distribution is very flat and wide, its value is inversely proportional to the width, so the marginal will again be small. Thus the marginal probability $P(Y|X)$ is a good criterion to see whether the prior is

consistent with the observed data, so it is also called *evidence*. A parameter like σ_w of the prior probability for a parameter w is called a *hyperparameter*, and the evidence is used for selection of prior probability, or hyperparameter tuning.

The same mechanism can also be used for selecting one of discrete candidates of probabilistic models $M_1, M_2, ...$ In this case the marginal probability for a model $P(M_i)$ can be used for *model selection*, then called *Bayesian criterion* for model selection.

1.6 Graphical Models and Other Bayesian Algorithms

So far we dealt with just two or three random variables, but in real life there are many states, observations, and parameters, some of which are directly or indirectly related. To make such dependency clear, graphical representations of random variables are useful. They are called *graphical models*, and the Bayes rule is used for estimation of the latent variables and parameters. Especially when a graphical model is represented by a *directed acyclic graph* (DAG), or equivalently, for an n-dim. variable vector X,

$$P(X) = \prod_{i=1}^{n} P(X_i|Pa_i), \tag{1.27}$$

where Pa_i denotes the parent variables of the variable X_i in the DAG, such a model is called a *Bayesian network*. For estimation of any missing variable in a Bayesian network, various *belief propagation* algorithms, the most famous one being the *message passing algorithm*, have been devised in recent years, and there are excellent textbooks to refer to when it becomes necessary for us to use one.

References

[1] Papoulis A (1991) *Random Variables, and Stochastic Process*. New York: McGraw-Hill.

[2] Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of Royal Society*, **53**, 370-418.

[3] Cox RT (1946) Probability, frequency and reasonable expectation. *American Journal of Physics*, **14**, 1-13.

[4] Shanon CE (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379-423, 623-656.

[5] Doucet A, de Freitas ND, Gordon N, eds. (2001) *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

[6] MacKay DJC (2003) *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press.