Proceedings of the 2018 European School of High-Energy Physics, Maratea, Italy, 20 June–3 July 2018
Edited by M. Mulders and C. Duhr, CERN-2019-006 (CERN, Geneva, 2019)

Practical Statistics for Particle Physicists

H. B. Prosper
Florida State University, Department of Physics, Tallahassee, USA

Abstract
These lectures cover the basic ideas of frequentist and Bayesian analysis and introduce the mathematical underpinnings of supervised machine learning. In order to focus on the essentials, we illustrate the ideas using two simple examples from particle physics.

Keywords
Statistics; lectures; data analysis method; statistical analysis; frequentist; Bayesian.

1 Introduction

Statistics and physics are similar in that each starts from sets of basic principles. They are similar also in the fact that physicists and statisticians from time to time engage in vigorous debate about the foundations of their respective disciplines. These disciplines, of course, also differ in significant ways. For example, physicists are forced, at some point, to bury the hatchet. Why? Because there is an ultimate judge of the correctness of a proposed principle, namely, *Nature*. If a principle yields results that contradict observations then the former does not apply to Nature and is, in that sense, wrong. For statisticians, alas, their many judges are other statisticians. Consequently, they are not compelled to reach, and for some basic questions have not reached, a consensus. Happily, however, for the typical applications in particle physics the debate and disagreements among statisticians can usually be ignored. But this is a poor excuse for dismissing these disagreements as even a modicum of understanding of them can avert hours of fruitless arguments that prove, ultimately, to be about intellectual taste and therefore cannot be adjudicated by appealing to a third party such as Nature. Therefore, while these lectures focus on the practical, we occasionally comment on some of these disagreements.

The remainder of the introduction, presents a birds eye view of statistical analysis. For detailed expositions on statistical analysis aimed at physicists, we recommend the books: [1–4]. For historical perspectives see [5, 6].

1.1 Samples

The result of an experiment is a sample of N data $X = x_1, x_2, \cdots, x_N$, which can be characterized with quantities called statistics[1]. A **statistic** is number that can be computed from the sample alone and known parameters. Here are a few well-known statistics:

the **sample moments**
$$x_r = \frac{1}{N}\sum_{i=1}^{N} x_i^r, \tag{1}$$

the **sample average**
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{2}$$

and the **sample variance**
$$s^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2. \tag{3}$$

[1] Statisticians tend to use upper case letters to denote random variables and lower case letters to denote actual values. We do not follow this convention.

0531-4283 – © CERN, 2019. Published under the Creative Common Attribution CC BY 4.0 Licence.

http://doi.org/10.23730/CYRSP-2019-006.261

The sample moments give detailed information about the sample, while the sample average and variance are measures of the center of the data and their spread. Statistics that characterize the data are called **descriptive statistics**. In these lectures, we shall encounter statistics that provide more sophisticated information about samples.

1.2 Populations

An infinitely large sample is called a **population**, which physicists usually refer to as an ensemble. Like other abstractions, populations can be studied mathematically and can be characterized with numbers, such as those listed below. (The symbol $E[*]$ means **ensemble average**, that is, the average over the population of the quantity within the brackets.)

Ensemble average	$E[x]$
Mean	μ
Error	$\epsilon = x - \mu$
Bias	$b = E[x] - \mu$
Variance	$V = E[(x - E[x])^2]$
Standard deviation	$\sigma = \sqrt{V}$
Mean square error	$\text{MSE} = E[(x - \mu)^2]$
Root MSE	$\text{RMS} = \sqrt{\text{MSE}}$

(4)

However, unlike the statistics of a sample, the numbers that characterize a population are abstractions. After all, no one has ever amassed an infinity of anything. In practice, a population is approximated by a large sample. Such “populations" are the basis of a statistical method called the bootstrap, in which various quantities can be approximated by treating the sample as if it were a population. Large, typically simulated, samples are used in physics analyses to assess, for example, the effect of systematic uncertainties or to confirm that an analysis method performs as claimed. In a simulated “population" some quantities can be computed exactly, for example the *error* associated with each element of the “population" because x and μ are known. Quantities such as bias, however, can only be approximated.

While it may not be possible to calculate a population quantity exactly, it often possible to relate one population quantity to another, which can sometimes provide useful insight. Take for example the mean square error (MSE), whose square root is called the root mean square (RMS)[2]. The MSE can be written as

$$\text{MSE} = V + b^2. \tag{5}$$

Exercise 1: Show this

This is an instructive result. Suppose, for example, that μ is the true Higgs boson mass and x is a measurement of it. If the MSE is used as a measure of the accuracy of the mass measurements, then the result in Eq. (6) shows that correcting a measurement of the mass for bias makes sense only if, on the average, the bias-corrected results yield a smaller MSE than that of the uncorrected result. Making a bias correction may not always be the correct thing to do if the goal is to arrive at mass measurements, which, on average, are as close to the true value of the mass in the MSE sense. Using simulations to study and understand the characteristics of a population is both useful and educational. It is good practice to do lots of simple simulations (sometimes called *toy* experiments) in order to develop an intuition about statistical quantities and the behavior of statistical procedures as well as to decide whether a particular manipulation of a measurement—e.g., a bias correction—makes sense.

[2]The RMS and standard deviation are sometimes used interchangeably. The two quantities are identical only if the bias is zero.

Another example of the ability (and utility) of mathematical analysis with respect to a population is the calculation of the bias in the variance of a sample. When we speak of "bias in a measurement x", say a measurement of the Higgs boson mass, we should remember that this phrasing is a shortcut. There is very likely an *error* in x, which in the real world is unknown. But, strictly speaking, *bias* does not apply to x, but rather to the ensemble to which x is presumed to belong. However, it would quickly become horribly pedantic not to use the shortcut "bias in x", so it is perfectly reasonable to use it so long as we remember what the phrase means. The ensemble average of the sample variance, Eq. (3), is given by

$$\begin{aligned} E[s^2] &= E[\overline{x^2}] - E[\bar{x}^2], \\ &= V - \frac{V}{N}. \end{aligned}$$

Exercise 2: Show this

and has a bias of $b = -V/N$. The result shows that the bias can be calculated exactly only if the variance V is known exactly.

1.3 Statistical Inference

The main goal of a theory of statistical inference is to use a sample to infer something about the associated population. We may wish to estimate (that is, measure) a parameter associated with the population, for example, the mean Higgs boson signal in the proton-proton to 4-lepton channel. Then, in order to make this estimate meaningful, we need to quantify its accuracy. Finally, we may wish to assess to what degree we can claim the signal is real and not an apparent signal caused by a fluctuation of the background. We shall consider each of these tasks using the two most commonly used theories of inference, frequentist and Bayesian. In both theories, the foundational concept is **probability**, albeit interpreted in two different ways:

- **Degree of belief** in, or assigned to, a proposition, e.g.,
  - *proposition*: it will rain in Maratea tomorrow
  - *probability*: $p = 5 \times 10^{-2}$
- **Relative frequency** of given outcomes in an infinite set of trials, e.g.,
  - *trial*: a proton-proton collision at the Large Hadron Collider (LHC)
  - *outcome*: creation of a Higgs boson
  - *probability*: $p = 5 \times 10^{-10}$

Since each theory of inference uses a different interpretation of probability, it is not surprising that the interpretation of their results differ. This can cause confusion, especially when both theories give numerically identical results. When data are plentiful, these interpretation typically do not affect how the results are subsequently used. The difficulties arise when sample sizes are small and when each approach can yield substantially different results. This is when intellectual taste becomes the main arbiter of which approach is considered the more reasonable.

The next two sections cover the application of frequentist and Bayesian theories of statistical analysis in particle physics using a simple real-word example, while the last section provides an introduction to supervised machine learning.

2 Frequentist Analysis

In 2014, the CMS Collaboration published its measurement of the properties of the Higgs boson in the 4-lepton final states [7]. We shall analyze the summary results of this analysis, namely, $N = 25$ observed

4-lepton events with a background estimate of $B \pm \delta B = 9.4 \pm 0.5$ events. The goal is to make statements about the mean Higgs boson event count s—that is, the signal, where $d = b + s$ is the mean event count and b is the mean background count. Although these data are very simple, they are sufficient to illustrate the essential ideas of frequentist analysis.

Whether the data are to be analyzed using a frequentist or Bayesian approach, the starting point is the same: the first task is constructing an accurate probability model for the mechanism that generates the data.

2.1 The Probability Model

Given the observed count $N = 25$ events, a particle physicist would immediately model the data generation mechanism with a Poisson distribution,

$$\text{Poisson}(n, d) = \frac{e^{-d} d^n}{n!},$$

because everyone knows that is the distribution for a counting experiment. If the data comprised M counts $N_m, m = 1, \cdots, M$ that are considered independent, the model would be a product of Poisson distributions. But, why is a Poisson appropriate? Let us start at the very beginning.

2.1.1 Bernoulli trial

A Bernoulli trial, named after the Swiss mathematician Jacob Bernoulli (1654 – 1705), is an experiment with only two possible outcomes: S, a success or F, a failure. Each collision between protons at the LHC is a Bernoulli trial in which either a Higgs boson is created (S) or is not (F). Here is a sequence of collisions results

$$F \quad F \quad S \quad F \quad F \quad F \quad F \quad S \quad F \quad \cdots$$

What is the probability of this sequence of results? There is *no* answer. Unless that is we are prepared to make assumptions, such as the following.

1. Let p be the probability of a success.
2. Let p be the same for every collision (trial).
3. Let S and F be *exhaustive* (the only possible outcomes) and *mutually exclusive* (one outcome precludes the occurrence of the other).

Assumption 3 implies that the probability of F is $1 - p$. Therefore, for a given sequence O of n proton-proton collisions, the probability $P(k|n, p, O)$ of exactly k successes and exactly $n - k$ failures is

$$P(k|n, p, O) = p^k (1 - p)^{n-k}. \tag{6}$$

The specific sequence O of successes and failures is unknown at the LHC. Whenever, we have a parameter that is either irrelevant or whose value is unknown, the rules of probability theory imply that the unknown can be eliminated from the problem by summing over all possible values of the unknown, here the orders of successes and failures O. This rule is called **marginalization** and is one of the most important procedures in probability calculations. Applied to our problem this yields,

$$P(k|n, p) = \sum_O P(k|n, p, O) = \sum_O p^k (1 - p)^{n-k}. \tag{7}$$

Notice that every term in Eq. (7) is identical and there are $\binom{n}{k}$ of them. Therefore,

$$P(k|n, p) = \binom{n}{k} p^k (1 - p)^{n-k}, \tag{8}$$

that is, we arrive at the **binomial distribution**, Binomial(k, n, p). If a is the mean number of successes in n trials, then

$$\begin{aligned} a &= \sum_{k=0}^{n} k \,\text{Binomial(k, n, p)}, \\ &= pn. \end{aligned} \tag{9}$$

Exercise 4: Show this

For the Higgs boson outcomes, $p \sim 10^{-10}$ and $n \gg 10^{12}$. Therefore, it is reasonable to consider the limit $p \to 0$ and $n \to \infty$, while keeping a constant. In this limit

$$\begin{aligned} \text{Binomial(k, n, p)} &\to e^{-a} a^k / k!, \\ &\equiv \text{Poisson}(k, a). \end{aligned} \tag{10}$$

Exercise 5: Show this

We conclude that a Poisson distribution is an appropriate model when the probability of individual events is extremely small. Indeed, the distribution can be derived from a stochastic model in which that assumption is made explicit. Therefore, it is indeed reasonable to take

$$p(n|s,b) = \text{Poisson}(n, s+b) = \frac{(s+b)^n e^{-(s+b)}}{n!}, \tag{11}$$

as the probability to obtain a count n given mean event count $s + b$.

We now turn to the probability model for the background data. In principle, the model should encode in detail how the background estimate was obtained. But, in order to keep matters as simple as possible, let us assume that the background estimate was obtained from an accurate Monte Carlo simulation, which yields a count m. The mean count in the simulation is kb, where k is a known scale factor that relates the mean count in the simulation to that in the signal region of the experiment that yielded N events. Therefore, the probability model for the background shall be taken to be

$$p(m|kb) = \text{Poisson}(m, kb). \tag{12}$$

Since the counts n and m are independent, the full model is

$$p(n, m|s, b) = \text{Poisson}(n, s+b)\text{Poisson}(m, kb). \tag{13}$$

2.2 The Likelihood Function

The **likelihood function** is the probability function—either a probability density function (pdf) if the random variables are continuous, or a probability mass function (pmf) if they are discrete—into which observations, that is, data have been inserted. Since the data are constants, the likelihood, $p(N, M|s, b)$ in our example, is a function of the parameters only. Sometimes, $p(N, M|s, b)$ is written as $L(s, b)$ to emphasize this point.

In this example, we are given $B \pm \delta B$, not M and k. But, we can infer M and k from B and δB using a plausible model, namely, that B and δB are M and $\sqrt{M}$ scaled down by k, that is,

$$B = M/k, \tag{14}$$

$$\delta B = \sqrt{M}/k. \tag{15}$$

Inverting these equations yields

$$M = (B/\delta B)^2 = 353.4, \tag{16}$$

$$k = B/\delta B^2 = 37.6. \tag{17}$$

Therefore, the likelihood for the count M is

$$(kb)^M e^{-kb}/\Gamma(M+1), \tag{18}$$

which we have written in a form that allows for non-integral values of M. Writing $D = N, M$, the full likelihood can be written as

$$p(D|s,b) = \frac{(s+b)^N e^{-(s+b)}}{N!} \frac{(kb)^M e^{-kb}}{\Gamma(M+1)}. \tag{19}$$

In a more realistic analysis, a probability model for the scale factor k would also be included. But, to keep things simple, we shall neglect the uncertainty in k.

Now that we have the likelihood function, several questions can be answered, including the following.

1. How is a parameter to be estimated?
2. How is its accuracy to be quantified?
3. How can an hypothesis be tested?
4. How is the statistical significance of a result to be quantified?

2.3 The Frequentist Principle

The goal of a frequentist analysis is to construct statements such that it can be guaranteed, *a priori*, that a fraction $f \geq p$ of them are true over an ensemble of similarly constructed statements. This stipulation is called the **frequentist principle** (FP) and was championed by the Polish statistician Jerzy Neyman [8]. The fraction f is called the **coverage probability**, or coverage for short, and p is called the **confidence level** (C.L.). An ensemble of statements that obey the frequentist principle is said to *cover*.

Points to Note

1. The FP applies to real ensembles [3], not just the virtual ones we simulate on a computer. Moreover, the ensembles can contain statements about different quantities. *Example*: all published measurements x, since the discovery of the electron in 1897, of the form $\theta \in [l(x), u(x)]$, where θ is a parameter of interest, that is, the parameter to be measured.
2. Coverage is an *objective* characteristic of ensembles of statements. However, in order to verify whether an ensemble of statements covers, we need to know which statements are true and which ones are false. Alas, since this information is generally not available in the real world there is no *operational* way to compute the coverage. The fact that we can do so in a simulation may give us confidence that the actual coverage of published statements is as the simulation reports, but does not prove that it is so.

Example

Consider an ensemble of different experiments, each with a different mean count θ, and each yielding a count N. Each experiment makes a single statement of the form

$$N + \sqrt{N} > \theta,$$

[3] Strictly speaking, we mean real samples because, as we have defined it, an ensemble is a synonym for a population, which by definition contains infinitely many elements

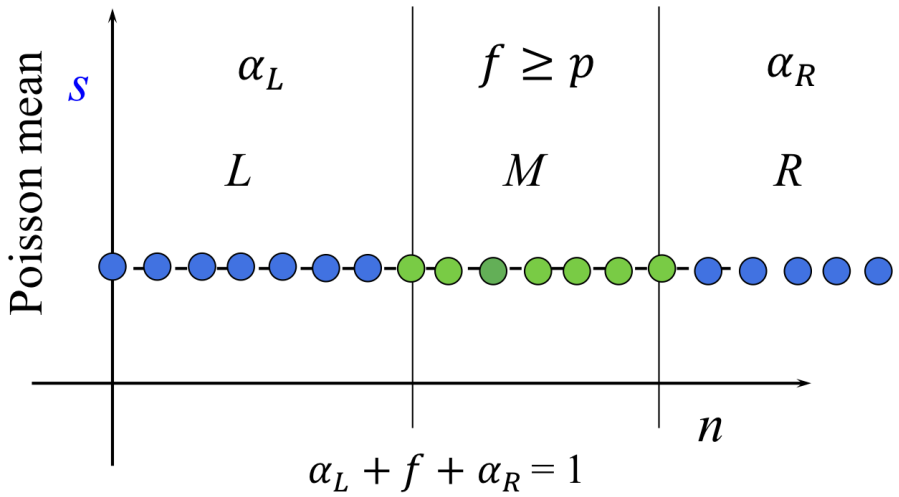

Fig. 1: Plotted is the tensor product of the parameter space, with parameter s, and the space of observations with potential observations n. For a given value of s, the observation space is partitioned into three disjoint intervals, labeled L, M, and R, such that the probability to observe a count n in M is $f \geq p$, where p = is the desired confidence level.

which is either true or false. If these were real experiments, we would not be able to determine which statements are true and which are false and, therefore, determine the coverage. Suppose that each mean count θ is randomly sampled from `uniform`(0, 10), with range [0, 10], and suppose that these means are known as would be the case in a simulation. Since the numbers are known, we can compute the coverage probability f.

Exercise 7: Compute the coverage of these statements; repeat the exercise using `uniform`(0, 1000)

In the next section, we discuss the important concept of the confidence interval, which is the classic exemplar of the frequentist principle.

2.4 Confidence Intervals

In 1937, Neyman [8] introduced the concept of the **confidence interval**, a way to quantify uncertainty in estimates that respects the frequentist principle. Confidence intervals are a concept best explained through an example. Consider an experiment that observes $n = N$ events with mean signal count s and no background. A confidence interval $[l(N), u(N)]$, with confidence level CL $= p$, permits a statement of the form

$$s \in [l(N), u(N)], \tag{20}$$

with the *a priori* guarantee that a fraction $f \geq p$ of statements will be true over an ensemble of such statements, not necessarily about the same quantity or the same kind of experiment. For simplicity, however, we shall consider experiments of the same kind, but which differ by their mean signal count s.

Consider Fig. 1, which shows the tensor product of the parameter space $\{s\}$ and the space of potential observations $\{N\}$ as well as the potential observations, represented by the dots, of an experiment with mean count s. The two vertical lines divide the space of observations into the three regions labeled L, M, and R. The region M is chosen so that the probability to obtain a count in that region is $f \geq p$, where p is the desired confidence level (CL). The probabilities to obtain a count in region L or region R are α_L and α_R, respectively. Since the three regions span the space of observations, $\alpha_L + f + \alpha_R = 1$.

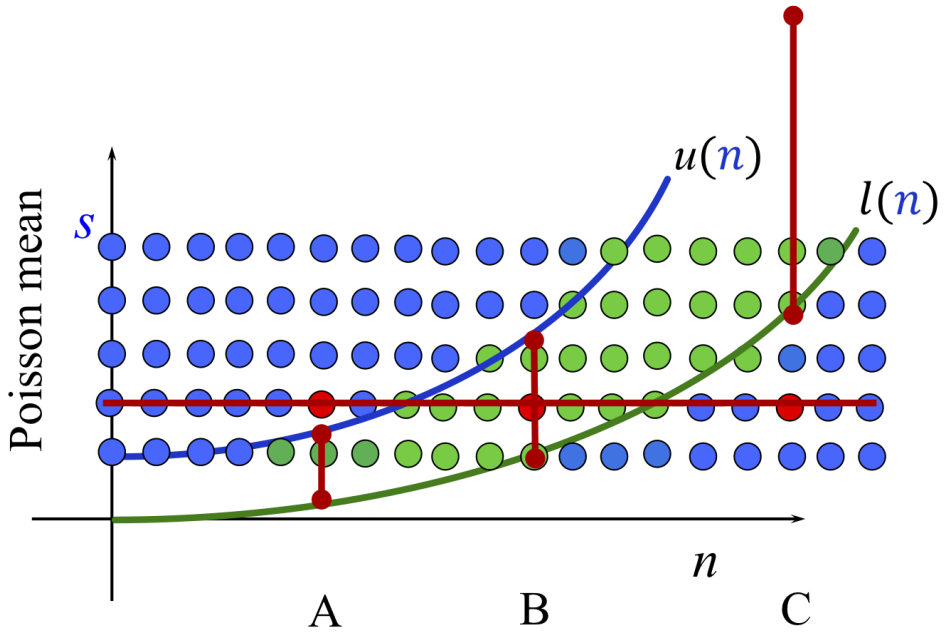

Fig. 2: The algorithm for defining region M (see Fig. 1), must be repeated for every value of s that is possible *a priori*. For the experiment whose mean s is represented by the thick horizontal line, the figure shows three possible outcomes, labeled A, B, and C, and their associated confidence intervals $[l(n), u(n)]$. Only outcomes, such as B, which lie within the region M of the experiment will yield intervals that bracket s. The probability to obtain such an interval is $f \geq p$, by construction.

For a given coverage f, the choice of region M is not unique and different methods have been suggested to define it. The first method was devised by Neyman [8], which we shall consider shortly. Another method was suggested by Feldman and Cousins [9]. We shall use that method to explain the general construction of confidence intervals.

Feldman-Cousins Method

In the Feldman-Cousins method, every potential count n is associated with a pair of numbers: a weight $p(n|s)\ /\ p(n|\hat{s})$, where $\hat{s} = n$ is the maximum likelihood estimate of s, together with the probability $p(n|s)$ to obtain the count n. The counts are placed in *descending* order of their weights. Starting with the first count in the ordered list, a set of counts $(n_{(1)}, n_{(2)}, \cdots)$ is accumulated one by one until their summed probabilities $f = \sum_{(i)} p(n_{(i)}|s) \geq p$. The symbol (i) denotes the ordinal value of a count in the ordered list. The set of counts $(n_{(1)}, n_{(2)}, \cdots)$ defines an interval in the space of observations whose lowest (leftmost) and highest (rightmost) counts n_L and n_R are given by $n_L = \min(n_{(1)}, n_{(2)}, \cdots)$ and $n_R = \max(n_{(1)}, n_{(2)}, \cdots)$, respectively. This construction (for this single parameter problem) guarantees that the probability to obtain a count within region M is $f \geq p$ [4].

There is, however, a snag with any algorithm to define M. The latter can only be defined if the mean count associated with an experiment is known. This may well be true within a simulation, but it is not so in the real world. Therefore, any algorithm for defining the region M must be repeated for every value of s that is considered possible *a priori*, as illustrated in Fig. 2. The repetition produces regions M_s, labeled by the mean count s, that define two curves, labeled $l(n)$ and $u(n)$, in the product space $\{s\} \otimes \{n\}$. For a given n, these curves define the confidence intervals $[l(n), u(n)]$. Over an ensemble

[4] We write $f \geq p$ rather than $f = p$ because, in general, for a discrete distribution it is not possible to satisfy the equality except at specific values of s.

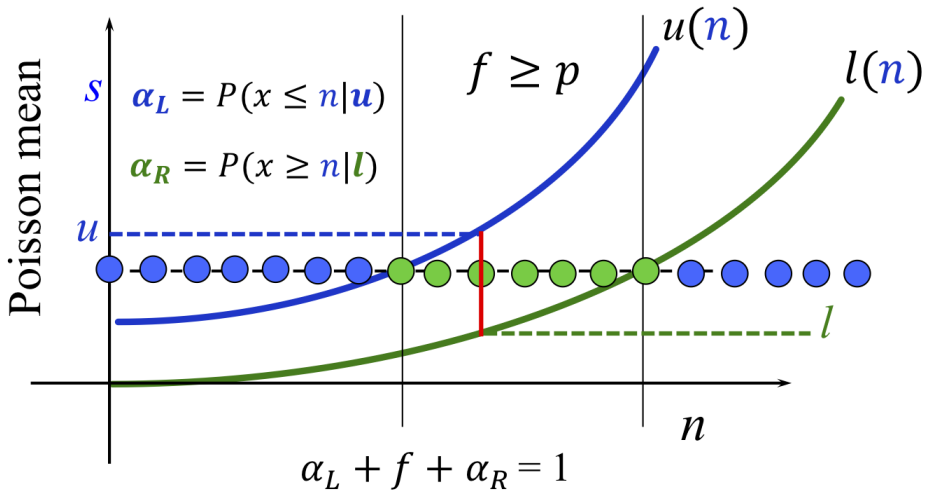

Fig. 3: The Neyman method. For every n, an interval $[l(n), u(n)]$ is computed by solving the equations in the plot. See text for details.

of experiments—and irrespective of their associated mean count s, the fraction of statements of the form $s \in [l(n),\ u(n)]$ that are true is $f \geq p$, by construction. To see this, consider again Fig. 2. It shows three possible outcomes for the experiment defined by the thick horizontal line together with the three possible confidence intervals (the vertical lines terminated with dots). If an observation lands in the region M for that experiment, the interval $[l(n), u(n)]$ will bracket the mean count s, as shown in the figure. If a count lands in region L, then the upper limit $u(n)$ will lie below s and, consequently, the interval $[l(n), u(n)]$ will exclude s. If n lands in region R, then the lower limit $l(n)$ will lie above s and the interval will exclude s. Therefore, the interval $[l(n), u(n)]$ will include s only if n lies in M, for which the probability is $f \geq p$. A procedure for constructing confidence intervals in this manner is called a **Neyman construction**.

Neyman Method

The algorithm described above requires that a region M be constructed for each value of s. An alternative algorithm was devised by Neyman in his 1937 paper and is illustrated in Fig. 3. For every n, the upper and lower limits are found by solving

$$P(x \leq n|u) = \alpha_L, \tag{21}$$
$$P(x \geq n|l) = \alpha_R. \tag{22}$$

Equation (21) yields a curve $u(n)$ for which the probability to obtain a count $x \leq n$, for a given s, is α_L, while Eq. (22) yields a curve $l(n)$ for which the probability to obtain a count $x \geq n$, for a given s, is α_R. These curves can also be made using the Neyman construction described above for the Feldman-Cousins method, but the solution using Eqs. (21) and (22) is computationally more efficient. Figure 4 shows the coverage probability over the parameter space for the Neyman intervals, in which we have chosen $\alpha_L = \alpha_R = (1 - p)/2$. This choice, the one made by Neyman, define **central confidence intervals**. As advertised, these confidence intervals satisfy the frequentist principle. Also shown is the coverage for intervals of the form $[N - \sqrt{N}, N + \sqrt{N}]$ and $[N - \sqrt{N}, N + \sqrt{N} + \exp(-N)]$. These intervals are *approximate* confidence intervals in that they do not satisfy the frequentist principle exactly. Notice, however, that for $s > 2.5$ the coverage of these intervals bounces around the $p = 0.683$ line. Therefore, over a large sample of experiments, with a distribution of Poisson means, it is plausible that the coverage could turn out to be close to the desired confidence level.

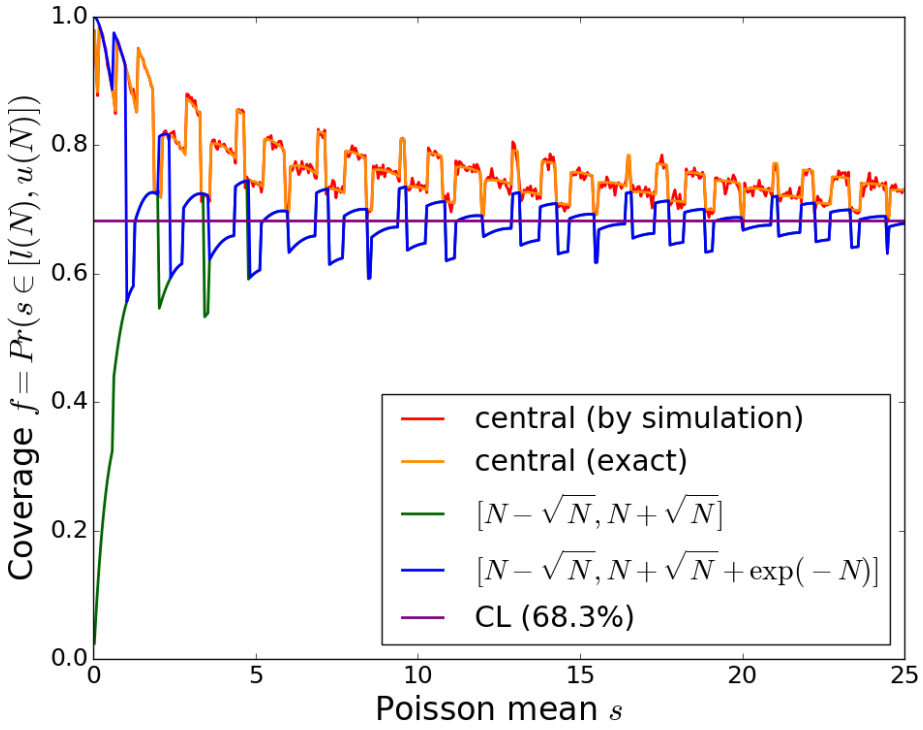

Fig. 4: Coverage probability f as a function of the Poisson mean s. As expected, the central intervals satisfy the frequentist principle, namely, $f \geq p$, where $p = 0.683$ is the confidence level. The coverage for two other sets of intervals are shown for which the frequentist principle is not satisfied.

A notable feature of Fig. 4 is the jaggedness of the coverage probabilities over the parameter space. The jaggedness is caused by the discreteness of the Poisson distribution. For a discrete distribution, coverage equal to the desired confidence level is possible only at specific values of s. Therefore, if we insist on the frequentist principle, $f \geq p$, the price to be paid is *over-coverage* in subsets of the parameter space.

2.5 The Profile Likelihood

The likelihood function,

$$p(D|s,b) = \frac{(s+b)^N e^{-(s+b)}}{N!} \frac{(kb)^M e^{-kb}}{\Gamma(M+1)}, \tag{23}$$

contains *two* parameters, the mean signal count s and mean background count b. However, the **parameter of interest** is the mean signal. The mean background count is needed to define the probability model, but inferences about it are not of interest. The parameter b is an example of a **nuisance parameter**. One way or another, we must rid a probability model of *all* nuisance parameters if we wish to make inferences about the parameter(s) of interest, here the mean Higgs boson signal count s. A widely accepted method for doing so is to convert the likelihood function into a function called the profile likelihood. But, before discussing this, we briefly describe the most common frequentist method to arrive at estimates of parameters.

Given the likelihood function $L(s,b) \equiv p(D|s,b)$, its parameters can be estimated by maximizing $L(s,b)$, or, equivalently, maximizing $\ln L(s,b)$, with respect to s and b,

$$\frac{\partial \ln p(D|s,b)}{\partial s} = 0 \quad \text{leading to } \hat{s} = N - B,$$

$$\frac{\partial \ln p(D|s,b)}{\partial b} = 0 \quad \text{leading to } \hat{b} = B,$$

as expected. Estimates found this way (first done by Karl Gauss and systematically developed by Fisher [10]) are called **maximum likelihood estimates** (MLE). This method generally leads to reasonable estimates, but, as is true of other procedures in statistical analysis, the method has its good and bad features, as noted below.

- *The Good*
  - Maximum likelihood estimates are *consistent*, that is, the RMS of estimates goes to zero as more and more data are included in the likelihood. This basically says that acquiring more data makes sense because the accuracy of results is expected to improve.
  - If an *unbiased* estimate of a parameter exists, the maximum likelihood procedure will find it.
  - Given the MLE for s, the MLE for any function $y = g(s)$ of s is $\hat{y} = g(\hat{s})$. This useful feature means that it possible to maximize the likelihood using any parameterization of it, say s, because, at the end, we can transform to the parameter of interest using $\hat{y} = g(\hat{s})$.
- *The Bad*
  - In general, MLEs are biased.

> **Exercise 7:** Show this
> Hint: Taylor expand $\hat{y} = g(s + \hat{s} - s)$ about s and consider its ensemble average.

- *The Ugly*
  - Most MLEs are biased, which, unfortunately, encourages the routine application of bias correction. But correcting for bias only makes sense if the RMS of an unbiased result is less than or equal to the RMS of a biased result. Recall that the RMS $= \sqrt{V + b^2}$, where V is the variance and b is the bias.

We now return to the profile likelihood. In order to make an inference about the signal, s, the 2-parameter model $L(s, b)$ must be reduced to one involving s only. In principle, this must be done while respecting the frequentist principle, that is, $f \geq p$, where f is the coverage probability of an ensemble of statements and p is the desired confidence level. In practice, all nuisance parameters are replaced by their MLEs conditional on given values of the parameters of interest. For the Higgs boson example, an estimate of b is found as a function of s, $\hat{b} = f(s)$, and b is replaced by $\hat{s}$ in $L(s, b)$. This leads to a function $L_p(s) = L(s, f(s))$ called the **profile likelihood**. For the likelihood in Eq. (23),

$$\hat{b} = f(s) = \frac{g + \sqrt{g^2 + 4(1+k)Ms}}{2(1+k)},$$
$$\text{where } g = N + M - (1+k)s. \tag{24}$$

Figure 5 shows a density plot of the likelihood $L(s, b)$ with the function $\hat{b} = f(s)$ superimposed. Notice that $\hat{b} = f(s)$ goes through the mode of $L(s, b)$, which occurs at $s = \hat{s} = N - B = 15.6$ events. Figure 6 shows the profile likelihood.

Replacing the (unknown) true value of b with an estimate thereof is clearly an approximation. Therefore, it should come as no surprise that inferences based on the profile likelihood are not guaranteed to be satisfy the frequentist principle exactly. However, it is found that for the typical applications in particle physics (as will be evident below), the procedures based on the profile likelihood work surprisingly well. Moreover, the use of the profile likelihood has a sound theoretical justification.

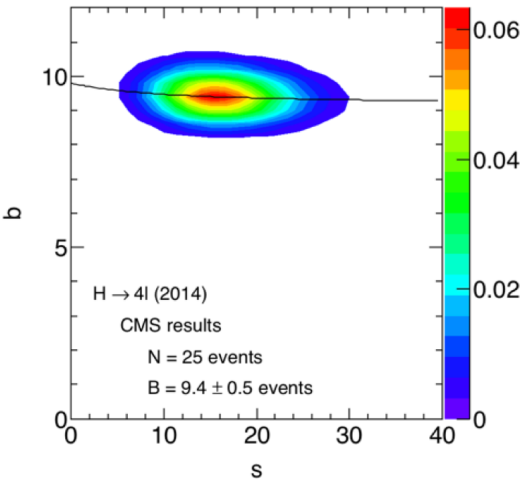

Fig. 5: The likelihood $L(s, b)$ and the graph of the function $\hat{b} = f(s)$.

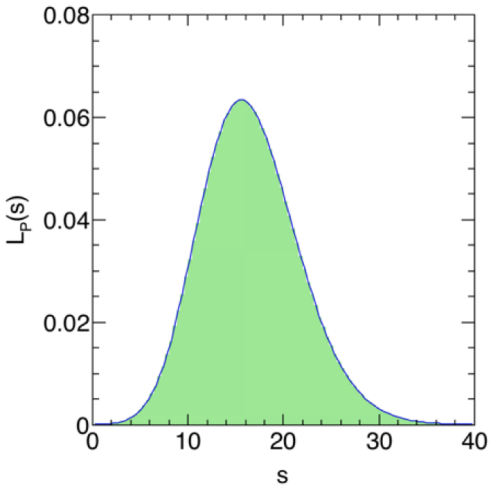

Fig. 6: The profile likelihood $L_p(s) \equiv L(s, f(s))$.

Consider the profile likelihood ratio

$$\lambda(s) = L_p(s)/L_p(\hat{s}), \tag{25}$$

where $\hat{s}$ is the MLE of s. Taylor expand the associated quantity

$$t(s) = -2 \ln \lambda(s) \tag{26}$$

about $\hat{s}$,

$$\begin{aligned} t(\hat{s} + s - \hat{s}) &= t(\hat{s}) + t'(\hat{s})(s - \hat{s}) \\ &\quad + t''(\hat{s})(s - \hat{s})^2/2 + \cdots \\ &\quad \approx (s - \hat{s})^2/2/\sigma^2 + \cdots , \end{aligned}$$

$$\text{where } \sigma^2 \approx 2/t''(\hat{s}). \tag{27}$$

The quadratic approximation is called the Wald approximation (1943) (see Cowan et al. [11]). If $\hat{s}$ does not occur on the boundary of the parameter space (in which case the derivative of t at $\hat{s}$ is zero), the sample is large enough (that is, when the density of $\hat{s}$ is approximately Gaussian$(\hat{s}, s, \sigma)$), and if s is the true value of the signal, then the density of $t(s)$ converges to a χ^2 density of one degree of freedom. The result, which is important because of its generality, is a special case of Wilks' theorem (1938) (Cowan et al. [11]).

Since $t(s) \approx \chi^2$, we can compute an approximate 68% confidence interval by solving

$$t(s) = -2 \ln \lambda(s) = 1, \tag{28}$$

for the lower and upper limits of the interval. Given $N = 25$ observed 4-lepton events, a background estimate of $B \pm \delta B = 9.4 \pm 0.5$, we can state that

$$s \in [10.9,\ 21.0] \quad @\ 68\%\ \text{C.L.} \tag{29}$$

Exercise 8: Verify this interval.

As noted, intervals constructed using the profile likelihood are not guaranteed to satisfy the frequentist principle. However, for applications in particle physics the coverage of these intervals is usually very good even for small amounts of data.

2.6 Hypothesis Tests

In the previous section, we concluded that $s \in [10.9, 21.0]$ @ 68% C.L. This result strongly suggests that a signal exists in the $N = 25$ 4-lepton events observed by CMS. But, a qualitative statement such as this is generally considered insufficient. The accepted practice is to perform an hypothesis test. Indeed, in particle physics, a discovery is declared only if a certain quantitative threshold has been reached in an hypothesis test.

An hypothesis test, in the frequentist approach, is a procedure for *rejecting* an hypothesis, which adheres to the following protocol.

1. Decide which hypothesis is to be *rejected*. This is called the **null hypothesis**. At the LHC, this is usually the background-only hypothesis.
2. Construct a function of the data called a **test statistic** with the property that large values of it would cast doubt on the veracity of the null hypothesis.
3. Choose a test statistic threshold above which we are inclined to reject the null. Do the experiment, compute the statistic, and reject the null if the threshold is breached.

We consider two related variants of this protocol, one by Fisher [10] and the other by Neyman, both developed in the 1930s. Fisher and Neyman disagreed strenuously about hypothesis testing, which suggests that the topic is rather more subtle than it seems. Fisher held that an hypothesis test required consideration of the null hypothesis only, while Neyman argued that a proper test required consideration of both a null as well as an alternative hypothesis. Physicists ignore these disagreements and see utility in an amalgam of the approaches of Fisher and Neyman. The is eminently sensible and pragmatic, whereas our quasi-religious adherence to a 5σ threshold before declaring a discovery is not always sensible.

We first illustrate Fisher's theory of hypothesis testing and follow with a description of Neyman's theory.

Fisher's Approach

We take the null hypothesis, which is denoted by H_0, to be the background-only model, that is, the Standard Model without a Higgs boson and compute a measure of the incompatibility of H_0 with the observations, called a **p-value**, defined by

$$\text{p-value}(x_0) = P(x > x_0|H_0), \qquad (30)$$

where x is a test statistic, designed so that large values indicate departure from the null hypothesis, and x_0 is the observed value of the statistic. Figure 7 shows the location of x_0. The p-value is the probability that x could have been higher than the x_0. Fisher argued that a sufficiently small p-value implies that either the null hypothesis is false or something rare has occurred. If the p-value is extremely small, say $\sim 3 \times 10^{-7}$, then of the two possibilities the response of the particle physicist is to reject the null hypothesis and declare that a discovery has been made. The p-value for our example, neglecting the uncertainty in the background estimate, is

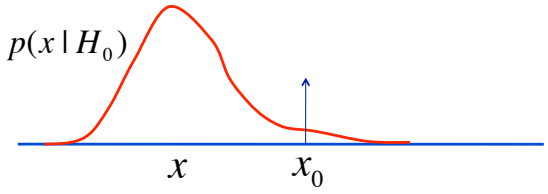

Fig. 7: The p-value is the tail-probability, $P(x > x_0|H_0)$, calculated from the probability density under the null hypothesis, H_0. Consequently, the probability density of the p-value under the null hypothesis is uniform$(0, 1)$.

$$\text{p-value} = \sum_{k=N}^{\infty} \text{Poisson}(k, 9.4) = 1.76 \times 10^{-5}, \text{ with } N = 25.$$

Since the p-value is a bit non-intuitive, it is conventional to map it to a Z**-value**, that is, the number of standard deviations the observation is *away from the null* if the distribution were a Gaussian. The Z-value can be computed using [5].

$$Z = \sqrt{2}\,\text{erf}^{-1}(1 - 2\text{p-value}). \qquad (31)$$

A p-value of 1.76×10^{-5} corresponds to a Z of 4.14σ. The Z-value can be calculated using the `Root` function

$$Z = \texttt{TMath::NormQuantile(1-p-value)}.$$

If the p-value is judged to be small enough, or the Z-value is large enough, the background-only hypothesis is rejected.

[5] $\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} \exp(-t^2)\, dt$ is the error funtion.

Neyman's Approach

As noted, Neyman insisted that a correct hypothesis test required *two* hypotheses to be considered, the null hypothesis H_0 and an alternative hypothesis H_1. This is illustrated in Fig. 8. The null is the same as before but the alternative hypothesis is the Standard Model with a Higgs boson, that is, the background plus signal hypothesis. Again, the statistic x is constructed so that large values would cast doubt on the validity of H_0. However, the Neyman test is specifically designed to respect the frequentist principle. A *fixed* probability α called the **significance (or size) of the test** is chosen, which corresponds to some threshold value x_α defined by

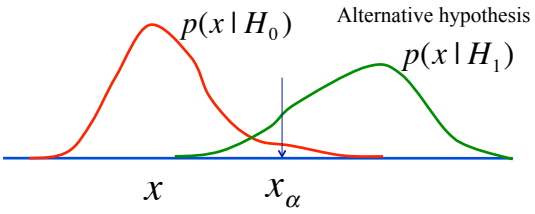

Fig. 8: Distribution of a test statistic x for two hypotheses, the null H_0 and the alternative H_1. In Neyman's approach to testing, $\alpha = P(x > x_\alpha|H_0)$ is a *fixed* probability called the significance of the test, which for a given class of experiments corresponds the threshold x_α. The hypothesis H_0 is rejected if $x > x_\alpha$.

$$\alpha = P(x > x_\alpha|H_0). \tag{32}$$

Should the observed value $x_0 > x_\alpha$, or equivalently, p-value$(x_0) < \alpha$, the hypothesis H_0 is rejected in favor of the alternative. By construction, a repeated application of this test will reject a fraction α of true null hypotheses. Since these are false rejections, we say that these are **Type I errors**. Neyman's test discards the p-value and reports only α and whether or not the null was rejected. However, in particle physics, in addition to reporting the results of the test, perhaps announcing a discovery, we also report the observed p-value. This makes sense because there is a more information in the p-value than merely reporting the fact that a null hypothesis was rejected at a significance level of α.

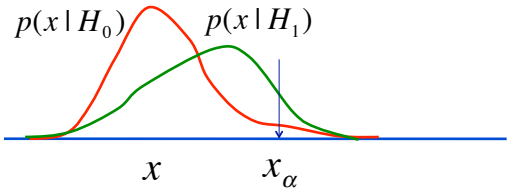

Fig. 9: See Fig. 8 for details. Unlike the case in Fig. 8, the two hypotheses H_0 and H_1 are not that different. It is then not clear whether it makes practical sense to reject H_0 when $x > x_\alpha$ only to replace it with an hypothesis H_1 that is not much better.

Given that Neyman's test requires an alternative hypothesis there is more that can be said than simply reporting the result of the test and the observed p-value. Figure 8 shows that we can also calculate

$$\beta = P(x \leq x_\alpha|H_1), \tag{33}$$

which is the relative frequency with which we reject true alternative hypotheses H_1. This mistake is called a **Type II error**. The quantity $1 - \beta$ is called the **power** of the test and is the relative frequency with which we would accept the true alternative hypotheses. The defining feature of the Neyman test is that, in accordance with the Neyman-Pearson lemma (see for example Ref. [2]), the power is maximized subject to the constraint that α is fixed. The Neyman-Pearson lemma asserts that given two simple hypotheses—that is, hypotheses in which all parameters have specified values—the optimal test statistic t for conducting an hypothesis test is the likelihood ratio $t = p(x|H_1)/p(x|H_0)$.

Maximizing the power seems like a reasonable procedure. Consider Fig. 9, which shows that the significance of the test in this figure is the same as that in Fig. 8. Therefore, the Type I error rates are identical. However, the Type II error rate is much greater in Fig. 9 than in Fig. 8 because the power of the test is considerably weaker in the former. Consequently, it is debatable whether rejecting the null is a wise course of action since the alternative hypothesis is not that much better. This insight was one source of Neyman's disagreement with Fisher. Neyman objected to the possibility that one might reject a null hypothesis regardless of whether it made sense to do so. He argued that the goal of hypothesis

testing is always one of deciding between competing hypotheses. Fisher's counter argument was that an alternative hypothesis may not be available, in which case we either give up or we have a method to test the only hypothesis that is available in order to decide whether it is worth keeping. In a Bayesian analysis an alternative hypothesis is also needed, in agreement with Neyman viewpoint, but is used in a way that neither he nor Fisher agreed with.

So far we have assumed that the hypotheses H_0 and H_1 are simple, that is, fully specified. Alas, most of the hypotheses that arise in realistic particle physics analyses are not of this kind. In the Higgs boson example, the probability models depend on a nuisance parameter for which only an estimate is available. Consequently, neither the background-only nor the background plus signal hypotheses are fully specified. Such hypotheses are examples of **compound hypotheses**. In the following, we illustrate how hypothesis testing proceeds in this case using the 4-lepton example.

Compound Hypotheses

In Sec. 2.5, we reviewed the standard way nuisance parameters are handled in a frequentist analysis, namely, their replacement by their conditional MLEs, thereby converting the likelihood function to the profile likelihood. In the 4-lepton example, this yielded the function $L_p(s) = L(s, f(s))$. The justification for this is that the statistic $t(s) = \ln \lambda(s)$, where $\lambda(s) = L_p(s)/L_p(\hat{s})$ and $\hat{s}$ is the MLE of s can be used to compute (approximate) confidence intervals in light of Wilks' theorem, which essentially states that $t(s) \approx \chi^2$. Therefore, the same statistic can also be used as a test statistic with the associated p-values calculated using the χ^2 density. Moreover, since, by definition, $Z = \sqrt{\chi^2}$, the p-value calculation can be sidestepped altogether. Using $N = 25$ and $s = 0$, we find $\sqrt{t(0)} = 4.13$, which is to be compared with $Z = 4.14$, the value found neglecting the ± 0.5 event uncertainty in the background.

In summary, the statistic $t(s)$ can be used to test null hypotheses as well as compute confidence intervals and, therefore, provides a unified way to deal with both tasks. If s is the true value of the mean signal, then the distribution of $t(s)$ under that hypothesis is a χ^2 density with one degree of freedom, $p(\chi^2|ndf = 1)$. Sometimes, however, it is necessary to consider $t(s)$ when the value of s in the argument differs from the value s, say s_0, which determines the density of $t(s)$. For example, suppose that a model of new physics predicts a mean count s_0 and an analysis is planned to test this model. We may be interested to know, for example, what value of $t(s)$ we might expect for a given amount of data. If $s = 0$, the goal may be to determine the average or median significance with which we may be able to reject the background-only hypothesis. Since the predicted signal s_0 differs from $s = 0$, the density of $t(s, \hat{s})$—where for clarity, the dependence on the estimate $\hat{s}$ is made explicit—will no longer be χ^2, but rather a non-central χ^2 density, $p(\chi^2|ndf = 1, nc)$ with non-centrality parameter nc, an approximate value for which is $nc = t(s, s_0)$; that is, it is the test statistic computed using an **Asimov**[6] data set [11] in which the "observed" count N is set equal to the true mean signal count, $s_0 + b$.

3 Bayesian Analysis

Bayesian analysis is merely applied probability theory with the following significant twist: a method is Bayesian if

- it is based on the degree of belief interpretation of probability and
- it uses Bayes' theorem

$$p(\theta, \omega|D) = \frac{p(D|\theta, \omega)\, \pi(\theta, \omega)}{p(D)}, \tag{34}$$

[6]The name of this special data set is inspired by the short story *Franchise* by Isaac Asimov describing a futuristic United States in which, rather than having everyone vote in a general election, a single (presumably representative) person is chosen to answer a series of questions whose answers are analyzed by an AI system. The AI system then decides the outcome of the election by determining what would have been the outcome had the general election been held!

where

$$
\begin{aligned}
D &= \text{ observed data,} \\
\theta &= \text{ parameters of interest,} \\
\omega &= \text{ nuisance parameters,} \\
p(D|\theta,\omega) &= \text{likelihood,} \\
p(\theta,\omega|D) &= \text{posterior density,} \\
\pi(\theta,\omega) &= \text{prior density,}
\end{aligned}
$$

for *all* inferences. The posterior density is the final result of a Bayesian analysis from which, if desired, various summaries can be extracted. The posterior density assigns a weight to every hypothesis about the values of the parameters of the probability model, which, in addition to the likelihood, also includes a function called the prior density or **prior** for short. The parameters can be discrete, continuous, or both, and nuisance parameters are eliminated by marginalization,

$$
\begin{aligned}
p(\theta|D) &= \int p(\theta,\omega|D)\, d\omega, \\
&\propto \int p(D|\theta,\omega)\, \pi(\theta,\omega)\, d\omega.
\end{aligned}
\tag{35}
$$

The prior $\pi(\theta,\omega)$ encodes whatever assumptions we make and information we have about the parameters θ and ω independently of the data D. A key feature of the Bayesian approach is recursion: the use of the posterior density $p(\theta,\omega|D)$ as the prior in a subsequent analysis.

These rules are simple, yet they yield an extremely powerful and general inference algorithm. However, particle physicists remain wedded to the frequentist approach because of the still widespread perception that the Bayesian algorithm is too subjective to be useful for scientific work. However, there is considerable published evidence to contrary, including in particle physics, witness the successful use of Bayesian analysis in the discovery of single top quark production at the Tevatron [16, 17] and searches for new physics at the LHC [18–20].

So, why do particle physicists, for the most part, remain skeptical about Bayesian analysis? For many, the Achilles heel of the Bayesian approach is the difficulty of specifying a believable prior over the parameter space of the likelihood function. In our example, in order to make an inference about the mean event count s using the data $N = 25$ events with a background of $B \pm \delta B = 9.4 \pm 0.5$ events, a prior density $\pi(s,b)$ must be constructed. Even after more than two centuries of effort, discussion, and argument, however, statisticians have failed to reach a consensus about how to do this in the general case. Nevertheless, Bayesian analysis is widely and successfully used, and used even within particle physics. This strongly suggests that we should refrain from overstating the difficulties. After all, physics is replete with approximations, both of a technical and conceptual nature. The same is true of statistical analysis. But, of course, this is no excuse for sloppiness. Rather it is a reminder not to make perfection the enemy of the good.

The particle physicists who have given this topic some thought seem to agree with the statisticians who argue that the following invariance property should hold for any prior, at least ideally,

$$
\pi_\phi(\phi)d\phi = \pi_\theta(\theta)d\theta,
\tag{36}
$$

where $\phi = f(\theta)$ is a one-to-one mapping of the parameter vector θ, e.g., $\theta = (s,b)$, to the new parameter vector ϕ and π_ϕ and π_θ are, in general, different functions of their arguments. If the above invariance holds, then the posterior density will likewise be reparametrization invariant in the same sense as the prior. Suppose we have a rule for creating a prior $\pi(*)$ and we apply this rule to create the density π_ϕ. The same rule is now used to create π_θ after which we transform from $\pi_\theta(\theta)d\theta$ to $\pi(\phi)d\phi$. Invariance with respect to the choice of parametrization demands that $\pi = \pi_\phi$. It surely ought not to matter whether

we parametrize the likelihood $p(D|s,b)$ in terms of s and b or in terms of s and $u = \sqrt{b}$. After all, the likelihood hasn't really changed, therefore, it would be odd if this "non-change" changed the posterior density. But, whether or not a change occurs depends on the nature of the prior, as the following example shows.

Consider the probability function $p(D|s) = \text{Poisson}(D|s)$, written in two different ways: $p(D|s) = \exp(-s)s^D/D!$ and $p(D|\sigma) = \exp(-\sigma^2)\sigma^{2D}/D!$, where $\sigma = \sqrt{s}$. In order to compute the posterior densities $p(s|D)$ and $p(\sigma|D)$ priors must be specified. The most widely used rule for doing so is: choose the prior to be flat, that is, uniform, e.g., $\pi(s) = 1$ and $\pi(\sigma) = 1$ in the parameter space. Notice that for an unbounded parameter space $\int \pi(s)\, ds = \int \pi(\sigma)\, d\sigma = \infty$. Yes, this has a bad look, but it is not necessarily a problem [12]! The posterior density in the s parametrization is $p(s|D) = \exp(-s)s^D/D!$, while it is $p(\sigma|D) = \exp(-\sigma^2)\sigma^{2D}/\Gamma(D+1/2)$ in the σ parametrization.

Now, if we transform $p(\sigma|D)d\sigma$ to $p'(s|D)ds$ the result is $p'(s|D) = \exp(-s)s^{D-1/2}/\Gamma(D+1/2)$, which clearly differs from $p(s|D)$. But, this is not surprising given that the flat prior is not reparametrization invariant. Some regard this as a serious problem, one that worsens as the dimensionality of the parameter space increases. Others point to the numerous successful uses of the uniform prior, even in problems with high dimensional parameter spaces, and accept the lack of invariance as a price worth paying in order to avoid the not inconsiderable effort of constructing an invariant prior.

A general method to create invariant priors was suggested by Jeffreys in the 1930s [15], which in the intervening years has received considerable mathematical validation through many different lines of reasoning (see, for example, [22]). The Jeffreys prior is given by

$$\pi(\theta) = \sqrt{\det I(\theta)}, \tag{37}$$
$$\text{where } I_{ij} = -E\left[\frac{\partial^2 \ln p(x|\theta)}{\partial\theta_i\,\partial\theta_j}\right] \text{ is the Fisher information matrix,}$$

and where the average is with respect to potential observations x sampled from the density $p(x|\theta)$. When the Jeffreys rule is applied to $p(x|\mu,\sigma) = \text{Gaussian}(x,\mu,\sigma)$ it yields

$$\pi(\mu,\sigma)d\mu\, d\sigma = \frac{d\mu\, d\sigma}{\sigma^2}. \tag{38}$$

Exercise 9: Show this

Ironically, the resulting posterior density was rejected by Jeffreys, and subsequently by statisticians because it yielded unsatisfactory inferences! The preferred prior for the Gaussian is

$$\pi(\mu,\sigma)d\mu\, d\sigma = \frac{d\mu\, d\sigma}{\sigma}, \tag{39}$$

because it leads to excellent results.

So, what is a confused physicist to make of this? One possibility is to reject the whole Bayesian omelette and stick to the frequentist gruel. It may be a tad thin for some, but it is at least relatively easy to make. The other is to dismiss the arguments that yield Eq. (37) in favor of reasoning that yields Eq. (39) (see, for example, [21]). Yet another way forward is to take seriously the many persuasive arguments that lead to Eq. (37) and try to understand what the reported failures of the Jeffreys prior for problems involving more than one parameter is telling us. Here is a hint of some understanding. Note that Eq. (39) can be written as

$$\begin{aligned}\pi(\mu,\sigma)d\mu\, d\sigma &= \sigma\,\left[\frac{d\mu\, d\sigma}{\sigma^2}\right], \\ &= \sigma_0 \exp(\ln\sigma/\sigma_0)\,\left[\frac{d\mu\, d\sigma}{\sigma^2}\right].\end{aligned} \tag{40}$$

This suggests, in the spirit of [22], that it is better to interpret the Jeffreys prior as simply an invariant measure on the parameter space of the associated likelihood function, one that assigns equal weight to every *probability density* labeled by θ. Assigning equal weight to every probability density is a reparametrization invariant procedure, while, as we saw above, assigning equal weight to every *parameter* is not. If this interpretation is accepted, then the prior density is actually given by

$$\pi(\theta) = g(\theta)\,\sqrt{\det I(\theta)}, \tag{41}$$

where $g(\theta)$ is a function that could assign non-equal weights to the probability densities, such as the term before the brackets in Eq. (40). That term is essentially the exponential of the entropy of the Gaussian density, which assigns a weight $\propto \sigma$ to every density indexed by μ, σ. What is missing is a convincing theoretical framework for choosing $g(\theta)$, a challenge that we leave to the reader.

For our example, we shall forego invariance in order to keep things simple and use a flat prior in both s and b. But, before delving back into the example, we review hypothesis testing in Bayesian analysis.

3.1 Model Selection

Hypothesis testing (also known as model selection) in Bayesian analysis requires the calculation of an appropriate posterior density or probability, as is true of all fully Bayesian calculations,

$$p(\theta, \omega, H|D) = \frac{p(D|\theta, \omega, H)\,\pi(\theta, \omega, H)}{p(D)}, \tag{42}$$

where we have explicitly included the index H to identify the different hypotheses. By marginalizing $p(\theta, \omega, H|D)$ with respect to all parameters except the ones that label the hypotheses or models, H, we arrive at

$$p(H|D) = \int p(\theta, \omega, H|D)\,d\theta\,d\omega, \tag{43}$$

that is, the probability of hypothesis H given observed data D. In principle, the parameters ω could also depend on H. For example, suppose that H labels different parton distribution function (PDF) models, say CT14, MMHT, and NNPDF, then ω would depend on the PDF model and should be written as ω_H. Like a Ph.D., it is usually convenient to arrive at the end-point, here the probability $p(H|D)$, in stages.

1. Factorize the prior, e.g.,

$$\begin{aligned}\pi(\theta, \omega_H, H) &= \pi(\theta, \omega_H|H)\,\pi(H),\\ &= \pi(\theta|\omega_H, H)\,\pi(\omega_H|H)\,\pi(H).\end{aligned} \tag{44}$$

   In many cases, we can assume that the parameters of interest θ are independent, *a priori*, of both the nuisance parameters ω_H as well as the model label H, in which case we can write, $\pi(\theta, \omega_H, H) = \pi(\theta)\,\pi(\omega_H|H)\,\pi(H)$.
2. Then, for each hypothesis, H, compute the function

$$p(D|H) = \int p(D|\theta, \omega_H, H)\,\pi(\theta, \omega_H|H)\,d\theta\,d\omega_H. \tag{45}$$

3. Then, compute the probability of each hypothesis,

$$p(H|D) = \frac{p(D|H)\,\pi(H)}{\sum_H p(D|H)\,\pi(H)}. \tag{46}$$

Clearly, in order to calculate the probabilities $p(H|D)$ it is necessary to specify the priors $\pi(\theta, \omega|H)$ and $\pi(H)$. With some effort, it is possible to arrive at an acceptable form for $\pi(\theta, \omega|H)$, however, it is highly unlikely that consensus could ever be reached on the prior $\pi(H)$. At best, we would have to make do with a convention. For example we could, by convention, assign equal probabilities to the two hypotheses H_0 and H_1, *a priori*, that is, $\pi(H_0) = \pi(H_1) = 0.5$. But, do we really believe that the Standard Model and the MSSM are equally probable models?

One way to sidestep the polemics of assigning $\pi(H)$ is to compare probabilities,

$$\frac{p(H_1|D)}{p(H_0|D)} = \left[\frac{p(D|H_1)}{p(D|H_0}\right] \frac{\pi(H_1)}{\pi(H_0)}, \tag{47}$$

but use only the term in brackets, called the global **Bayes factor**, B_{10}, as a way to compare hypotheses. The Bayes factor is the factor by which the relative probabilities of two hypotheses *changes* as a result of incorporating the data, D. The word global indicates that we have marginalized over all the parameters of the two models. The *local* Bayes factor, $B_{10}(\theta)$ is defined by

$$B_{10}(\theta) = \frac{p(D|\theta, H_1)}{p(D|H_0)}, \tag{48}$$

where,

$$p(D|\theta, H_1) \equiv \int p(D|\theta, \omega_{H_1}, H_1)\, \pi(\omega_{H_1}|H_1)\, d\omega_{H_1}, \tag{49}$$

are the **marginal** or integrated likelihoods in which we have assumed the *a priori* independence of θ and ω_{H_1}. We have further assumed that the marginal likelihood H_0 is independent of θ, which is a very common situation. For example, θ could be the expected signal count s, while $\omega_{H_1} = \omega$ could be the expected background b. In this case, the hypothesis H_0 is a special case of H_1, namely, it is the same as H_1 with $s = 0$. An hypothesis that is a special case of another is said to be **nested** within the more general hypothesis. All this will become clearer when we work through the Bayesian analysis of the 4-lepton data.

There is a notational subtlety that may be missed: because of the way we have defined $p(D|\theta, H)$, we need to multiply $p(D|\theta, H)$ by the prior $\pi(\theta)$ and then integrate with respect to θ in order to calculate $p(D|H)$.

3.2 Bayesian Analysis of 4-lepton Data

In this section, we shall

1. compute the posterior density $p(s|D)$,
2. compute a 68% credible interval $[l(D), u(D)]$, and
3. compute the global Bayes factor $B_{10} = p(D|H_1)/p(D|H_0)$,

as a way to illustrate a Bayesian analysis of the 4-lepton data.

Probability model

The likelihood is the same as that used in the frequentist analysis, namely, Eq. (23). However, the likelihood is only part of the model; we also need a prior $\pi(s, b)$ that encodes what we *know*, or *assume*, about the mean background and signal independently of the observations D. How exactly that should be done remains an active area of debate and research. Below, we shall take the easy way out!

One point that should be noted is that the prior $\pi(s, b)$ can be factorized in two ways,

$$\pi(s, b) = \pi(s|b)\, \pi(b),$$

$$= \pi(b|s)\,\pi(s). \tag{50}$$

It is worth noting because $\pi(s,b)$ is routinely written as $\pi(s,b) = \pi(s)\pi(b)$, which is not true, in general. The *a priori* independence of s and b is an assumption, one that we shall make. What do we know about s and b? We know that s and b are ≥ 0. We also know the probability function and how s and b enter it. Given this information, there are well founded methods to construct $\pi(s,b)$. However, for simplicity, for b, we shall use the improper prior $\pi(b) = k$, where k is the scale factor in the likelihood $p(D|s,b)$, and either the improper prior $\pi(s) = 1$, or the proper prior $\pi(s) = \delta(s - 15.6)$. An improper prior is one that integrates to infinity, which as noted above is not necessarily problematic [12].

Marginal likelihood

Having completed the probability model, the rest of the Bayesian analysis proceeds in a routine manner. First, it is convenient to eliminate the nuisance parameter b, using the improper prior $\pi(b) = k$,

$$\begin{aligned} p(D|s,H_1) &= \int_0^\infty p(D|s,b)\,\pi(b)\,db, \\ &= \frac{1}{M}(1-x)^2 \sum_{r=0}^{N} \text{Beta}(x, r+1, M)\,\text{Poisson}(N-r|, s), \end{aligned} \tag{51}$$

where $x = 1/(1+k)$,

Exercise 10: Show this

and thereby arrive at the marginal likelihood $p(D|s,H_1)$. The symbol H_1 has been introduced to represent the hypothesis that the signal is non-zero.

Posterior density

Given the marginal likelihood $p(D|s,H_1)$ and $\pi(s)$ we can compute the posterior density,

$$p(s|D,H_1) = p(D|s,H_1)\,\pi(s)/p(D|H_1), \tag{52}$$

where,

$$p(D|H_1) = \int_0^\infty p(D|s,H_1)\,\pi(s)\,ds.$$

Setting $\pi(s) = 1$ yields,

$$p(s|D,H_1) = \frac{\sum_{r=0}^{N} \text{Beta}(x, r+1, M)\,\text{Poisson}(N-r|s)}{\sum_{r=0}^{N} \text{Beta}(x, r+1, M)}. \tag{53}$$

Exercise 11: Derive an expression for $p(s|D,H_1)$ assuming $\pi(s) = \text{Gamma}(qs, 1, U+1)$ where q and U are known constants.

The posterior density $p(s|D,H_1)$ completes the inference about the mean signal s. In principle, we could stop there, but, in practice, summaries of the posterior density are furnished, such as a **credible interval**, the analog of a confidence interval. But, like confidence intervals, credible intervals, $[l(D), u(D)]$ with credible level p, defined by

$$\int_{l(D)}^{u(D)} p(s|D,H_1)\,ds = p \tag{54}$$

are not unique. The analog of Neyman's central interval is the central credible interval defined by

$$\int_0^{l(D)} p(s|D,H_1)\,ds = (1-p)/2,$$

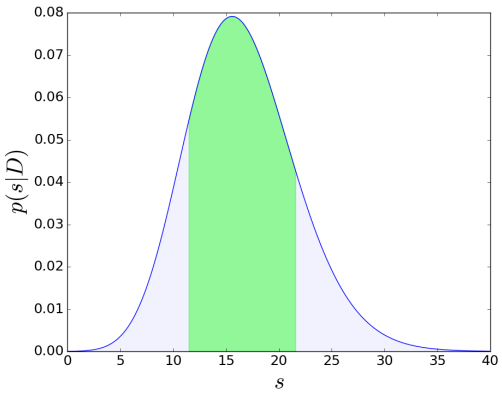

Fig. 10: Posterior density for 4-lepton data. The shaded area is the 68% central credible interval.

$$\int_{u(D)}^{\infty} p(s|D, H_1)\, ds = (1 - p)/2. \tag{55}$$

For the 4-lepton data this leads to the central credible interval $[11.5, 21.7]$ for s with $p = 0.683$, which is shown in Fig. 10. The statement $s \in [11.5, 21.7]$ at 68% C.L. means there is a 68% probability that s lies in the specified interval. Unlike the analogous frequentist statement, this one is about this particular interval and the 68% is a degree of belief, not a relative frequency. Statements of this form do, of course, have a coverage probability. However, *a priori*, there is no reason why the coverage probability of credible intervals should satisfy the frequentist principle. In practice, it is found that credible intervals with appropriately chosen priors can moonlight as approximate confidence intervals. But when this happens if does not mean that their interpretations somehow merge, it simply means that a misinterpretation of the intervals is likely to be benign.

Bayes factor

We noted above that

$$p(D|H_1) = \int_0^{\infty} p(D|s, H_1)\, \pi(s)\, ds.$$

Furthermore, $p(D|H_1) < \infty$ even with the improper prior $\pi(s) = 1$. However, another *arbitrary* constant besides unity could have been chosen, for example, $\pi(s) = C$. That constant would not have altered the posterior density $p(s|D, H_1)$ and therefore choosing $C = 1$ as a matter of convenience was fine. However, here we wish to compute the global Bayes factor $B_{10} = p(D|H_1) \;/\; p(D|H_0)$. The background-only hypothesis, H_0, is nested in H_1 and has marginal likelihood $p(D|H_0) \equiv p(D|0, H_1)$. Since the constant k in the background prior $\pi(b) = k$ scales both $p(D|H_1)$ and $p(D|H_0)$ the constant cancels and no issue arises from using an improper background prior. However, since for H_1 $\pi(s) = C$ and the parameter s appears only in the calculation of $p(D|H_1)$, the Bayes factor is scaled by the arbitrary constant C. Consequently, the Bayes factor can be assigned any value merely by choosing an appropriate value for C. This is clearly unsatisfactory. The upshot is that while improper priors may yield reasonable results for the posterior density $p(s|D, H_1)$, albeit ones that are not reparametrization invariant, that is not the case for Bayes factors. To arrive at a satisfactory Bayes factor, a proper prior must be used. The simplest such prior is, for example, $\pi(s) = \delta(s - \hat{s})$, where $\hat{s} = N - B = 15.6$ events. With this prior, the Bayes factor is

$$B_{10} = \frac{p(D|H_1)}{p(D|H_0)} = 4967.$$

We conclude that the 4-lepton observations increase the probability of hypothesis $s = 15.6$ events relative to the probability of the hypothesis $s = 0$ by ≈ 5000. In order to avoid large numbers, the Bayes factor can be mapped into a measure akin to the frequentist "n-sigma",

$$Z = \sqrt{2 \ln B_{10}}, \tag{56}$$

which gives $Z = 4.13$.

The Bayesian and frequentist results are approximately the same, which is typically the case when the data are sufficient. This is because the influence of the prior is smaller than when the data are sparse.

4 Supervised Machine Learning

The project of creating artificial beings that mimicked some characteristics of humans has been a dream of visionaries for millennia. But, during the Second World War, dreams gave way to a desperate focus on matters of life and death when latter day visionaries sought to create algorithms that could solve difficult problems such as cracking military codes in real-time. After the war, the pursuit of artificially intelligent agents was revived. In 1950, the great English mathematician Alan Turing, whose genius helped save millions of lives and shortened the most calamitous war in history, proposed an operational definition of such an agent, a test now known as the *Turing test* [23]. The test cuts to the chase regarding what it means to be intelligent: if it is impossible to tell whether one is conversing with a person or a machine and you are in fact conversing with a machine then the latter is intelligent. In the decades following the publication of the Turing test, progress towards creating such agents was slow, in part because the required conceptual breakthroughs were lacking and in part because the available computing power was severely limited.

However, an enormous change has occurred during the last decade or so, driven in part by algorithmic breakthroughs, but mostly by the exponential growth in the size of data sets and the available computing power. In just a few years, the field of *machine learning*, that is, the use of computer-based algorithms to construct useful models of data, has gone from research lab to everyday commercial applications. To be sure, there are many things humans do that seem far beyond current machine learning capabilities. It is still the case that we are unable to replicate a young child's ability to intuit the fact that the noises she hears from the people around her have meaning. Nor can we replicate the extraordinary human ability to be "trained" on a relatively small number of instances of, say, pictures of the Golden Gate bridge, and yet be able to identify the Golden Gate in other pictures of the bridge taken from perspectives that may never have been seen before. Nevertheless, impressive progress has been made recently. Arguably, the most notable is the breakthrough by the Google subsidiary *DeepMind* in creating an agent that taught itself to play to superhuman levels the ancient Chinese game of Go, as well as Chess and Shogi (Japanese chess) *tabula rasa*. These self-teaching feats were achieved in a mere 24 hours [24]!

Our purpose here is considerably more modest; it is to emphasize something that can easily get lost in the hype, namely, that these systems are, for the most part, "simply" highly non-linear high-dimensional parameter space functions that provide mappings from one space to another. The breakthrough has been the ability to fit these enormously complicated functions on practical timescales. In order to avoid complications that merely obfuscate, we consider a simplified version of the following problem: separating Higgs boson events in which the Higgs boson is produced via vector boson fusion (VBF) from events in which the Higgs boson is created via gluon gluon fusion (ggF). But, first, we give an overview of a few key ideas of machine learning.

Most machine learning algorithms fall into five broad categories:

1. supervised learning,
2. semi-supervised learning,

3. unsupervised learning (i.e., pattern detection)
4. reinforcement learning, and
5. generative learning.

The simplest category of algorithm is supervised learning in which the data for fitting models, i.e., training them, consist of labeled objects. If the labels define the class to which objects belong, for example, -1, or 0, for gluon gluon fusion events and $+1$ for vector boson fusion events, then, as shown below, the resulting function will be a *classifier*. If the labels form a continuous set, then the resulting function will be a *regression function* (sometimes called a "regressor"). For example, suppose the objects are jets characterized by their transverse momentum p_T and pseudo-rapidity η and possibly other detailed characteristics, such as the electromagnetic fraction, while the labels are the true jet transverse momenta. The regressor will be a correction function that maps the jet characteristics to an approximation of the true jet p_T. Our example will be a simple VBF/ggF classifier.

4.1 A Bird's Eye View of Supervised Learning

Supervised machine learning can be construed as a game in which winning means picking the best function (or functions) from a function space. The game includes three elements:

1. a function space $\mathcal{F} = \{f(x, w)\}$ containing parametrized functions $f(x, w)$, where x are object characteristics—*features* in machine learning jargon—and w are the parameters;
2. a loss function $L(y, f)$, which measures the cost of making a bad function choice, and where y are labels associated with the features x, and
3. a constraint $C(w)$ that places some restriction on the choices.

The best function $f(x, w*)$ is found by minimizing the constrained *empirical risk*,

$$R(f) = \sum_{i=1}^{K} L(y_i, f_i) + C(w), \text{ where } f_i = f(x_i, w), \tag{57}$$

with respect to the choice of function f, which in practice means with respect to the parameters w.

Minimization via Gradient Descent

A loss function, through the empirical risk, defines a "landscape" in the space of parameters, or equivalently in the space of functions. The goal is to find the lowest point in that landscape, usually by moving in the direction of the local negative gradient,

$$w_j \leftarrow w_j - \rho \frac{\partial R}{\partial w_j}, \quad j = 1, \cdots J, \tag{58}$$

where ρ is called the learning rate and J is the dimensionality of the parameter space, which, in some recent commercial applications can be in the millions. As is, the algorithm in Eq. (58) would fail miserably because of the complexity of the landscape and the possibility that the minimizer could get stuck in a local minimum or diverge away from the minimum because of the instability caused by a saddle point. To alleviate this problem, the standard approach is to replace the exact derivatives $\partial R/\partial w_j$ by *noisy* estimates thereof. This is usually achieved by replacing R by an approximation that uses a small subset—that is, *batch*—of the training data in the sum that defines R. Typically, a new batch is used at every step of the minimization algorithm. This minimization algorithm is called *stochastic gradient descent*, of which there are many variations. The addition of noise increases the chance that the minimizer will escape from an unfavorable location in the parameter space.

To Infinity and Beyond

It is intuitively clear that a successful minimization of the empirical risk, Eq. (57), will yield a solution $f(x, w*)$ that is as close as possible to the labels, or **targets**, y. But, in mathematics, as in physics, we can gain a a clearer understanding of a construct by taking a suitable limit of it. To that end, consider the limit of $R(f)/K$, that is, the *average loss*, as $K \to \infty$. Writing the average loss in that limit as E, and assuming that the effect of the constraint goes to zero in that limit, we can write

$$\begin{aligned} E[f] &= \int dx \int dy\, L(y, f)\, p(y, x), \\ &= \int dx\, p(x) \left[\int dy\, L(y, f)\, p(y|x) \right], \end{aligned} \tag{59}$$

where we have used $p(y|x) = p(y, x)/p(x)$. The function $p(y, x)$ is the (typically unknown) joint probability density of the targets and features (y, x). Whether the features x represent an event, a jet, an image, or piece of writing, and y represents useful known data about each instance of x, all the information about the mapping from x to y is contained in the joint probability density $p(y, x)$. This is an important point because the failures of machine learning are almost always due to an object with known characteristics x', but unknown label y', not being a member of the population $\{(y, x)\}$ that defines $p(y, x)$. If an agent is trained on a million images of dogs and cats, it is not surprising that it will classify a horse as either a dog or a cat because the probability density $p(y, x)$ does not encompass images of horses. The point is that the function $f(x, w)$ will do what it is designed to do. But, what exactly is $f(x, w)$ designed to do? To answer this question concretely, let us consider the minimization of Eq. (59) with the widely used **quadratic loss**,

$$L(y, f) = (y - f)^2. \tag{60}$$

If we change the function f by an arbitrary amount δf this induces a change

$$\delta E = 2 \int dx\, p(x)\, \delta f(x, w) \left[\int dy\, (y - f)\, p(y|x) \right], \tag{61}$$

in the mean loss E, which, in general, is not zero. If, however, the function $f(x, w)$ is sufficiently flexible, it will be possible to reach the minimum of E, where $\delta E = 0$. But, we want this to hold for all variations δf—because these variations are, after all, arbitrary—and for all values of x in order that the function f not fail—that is, perform poorly—for some subset of the space of features. This can be assured provided that the quantity in brackets in Eq. (61) is zero, that is, if

$$\boxed{f(x, w*) = \textstyle\int y\, p(y|x)\, dy.} \tag{62}$$

Equation (62) is an important result because it tells us precisely what the function $f(x, w*)$ approximates. If one uses the quadratic loss, then the function $f(x, w*)$ approximates the conditional average of the targets. This result was first derived in the context of neural networks [25–27], however, the result holds irrespective of the details of the function $f(x, w)$. In particular, the function does not have to be a neural network. The result holds provided that

1. we use sufficient training data $T = \{(y, x)\}$,
2. we use a sufficiently flexible function $f(x, w)$ and
3. we use an appropriate loss function.

Moreover, if we choose targets in the discrete set $y \in \{0, 1\}$, the general result reduces to

$$\boxed{f(x, w*) = p(y = 1|x)} \tag{63}$$

We conclude that if we minimize the average quadratic loss using training data in which one class of objects is labeled with $y = 0$ and the other with $y = 1$, the function $f(x, w*)$ approximates the probability that the object with features x belongs to the class labeled with $y = 1$; that is, $f(x, w*)$ is a classifier that approximates the class probability. From Bayes theorem, this class probability, $p(1|x)$, can be written as

$$p(1|x) = \frac{p(x|1)\,p(1)}{p(x|1)\,p(1) + p(x|0)\,p(0)}, \tag{64}$$

where $p(1)$ and $p(0)$ are the prior probabilities associated with the two classes. Typically, one trains with $p(1) = p(0)$, in which case $p(1|x)$ is referred to as a discriminant, $D(x)$, and is given by

$$D(x) = \frac{p(x|1)}{p(x|1) + p(x|0)}. \tag{65}$$

Boosted Decision Trees

Boosted decision trees (BDT) [28] are, currently, the most popular machine learning method in particle physics; and for good reason. They perform well, they are faster to train than neural networks, they are insensitive to poorly performing variables, and they are resistant to overfitting. In view of their widespread use, it is worth taking the time to understand exactly what this machine learning model entails. We shall highlight key features of BDTs using a simple example in which we seek to separate Higgs boson events produced via vector boson fusion (VBF) from gluon gluon fusion (ggF) produced events. In this section, we first discuss decision trees (DT) and then the notion of boosting, that is, enhancing the performance of a machine learning model by averaging over many models.

A decision tree is a nested sequence of *if then else* statements, which can also be viewed as a histogram whose bins are created recursively through *binary partitioning*. The VBF/ggF example uses two discriminating variables (features) $|\Delta\eta|_{jj}$ and m_{jj}, the absolute pseudo-rapidity difference between the two most forward (i.e., largest rapidity) jets in the event and the associated di-jet mass, respectively. Figure 11 shows two representations of a decision tree for our VBF/ggF discrimination example.

At face value, decision trees do not seem to fit into the mathematical ideas about loss functions discussed above. In particular, it is far from clear what loss function, if any, is being minimized when a decision tree is grown. However, all successful uses of decision trees entail averaging over many of them. As we shall see, it is the averaging that provides the connection to a loss function. Averaging also mitigates a serious problem with decision trees, namely, their instability. Even minor changes to the training data can radically alter the structure of a tree.

The first successful averaging algorithm, called `AdaBoost`, was published by AT&T researchers Freund and Schapire in 1997 [29] who showed that it was possible to create high performance classifiers by averaging ones (called **weak learners**) that perform only marginally better than classification via a coin toss. The algorithm builds a classifier using training data labeled by the discrete labels $y = -1$ or $y = +1$. In the VBF/ggF example below, $y = -1$ is assigned to ggF events and $+1$ is assigned to VBF events. The algorithm, for N training events and K decision trees, proceeds as follows:

1. **initialize** event weights $\omega_{1,n} = 1/N$, $n = 1, \cdots, N$
2. **repeat for** $k \in 1, \cdots K$
   (a) fit a tree $f_k(x)$ that returns either -1 or $+1$, using the current event weights $\{w_{k,n}\}$
   (b) compute error rate $\epsilon_k = \sum_{n=1}^{N} \omega_{k,n} I[-y_n f_k(x_n)]$, $I(z) = 1$ if $z > 1$, 0 otherwise
   (c) compute coefficient $\alpha_k = \frac{1}{2} \ln[(1 - \epsilon_k)/\epsilon_k]$
   (d) update weights $w_{k+1,n} = w_{k,n} \exp(-\alpha_k y_n f_k(x_n))/Z_k$,
   where $Z_k = \sum_{n=1}^{N} \omega_{k,n} \exp(-\alpha_k y_n f_k(x_n))$
3. classifier $f(x) = \sum_{k=1}^{K} \alpha_k\, f_k(x)$

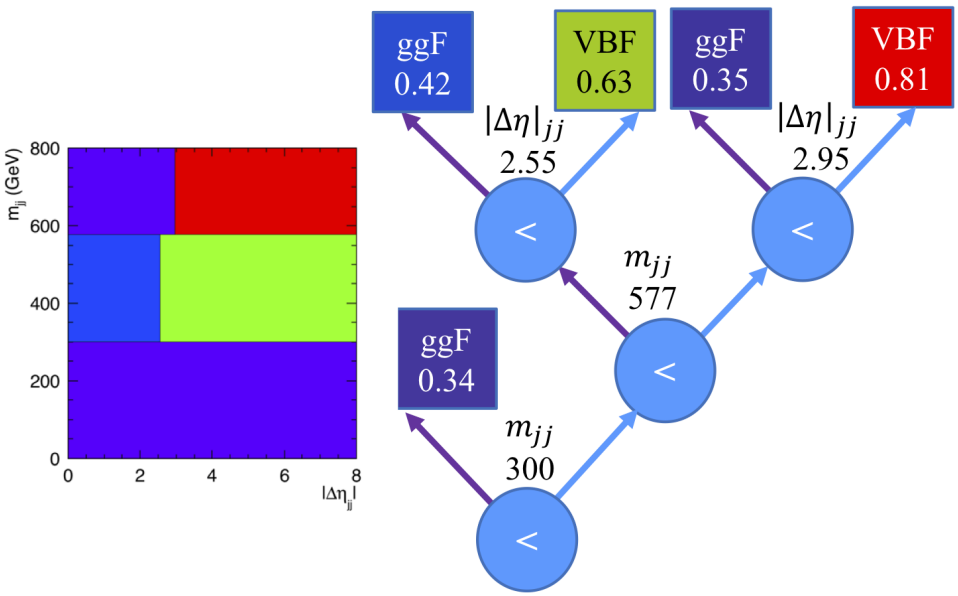

Fig. 11: Two representations of a decision tree to separate VBF from ggF events based on the variables $|\Delta\eta|_{jj}$ and m_{jj}. On the right, the decision tree is represented as a branching structure in which the circles, called *nodes*, represent *if then else* decisions, that is, *binary* decisions. The boxes terminate the tree and are referred to, appropriately, as *leaves*. On the left, the decision tree is represented as a 2D histogram in which the bins, which correspond to the leaves, have been defined by recursive binary partitioning. The bin boundaries, that is, the binary partitions, correspond to the decisions. At a given node, the left branch is taken if $x < x_{\text{cut}}$ otherwise the right branch is taken; x_{cut} is an optimal cut on the variable $x \in \{|\Delta\eta|_{jj}, m_{jj}\}$. The numbers within the leaves are the VBF purity $p = S/(S+B)$, where S and B are the VBF and ggF event counts in a given bin, that is, leaf.

In step 2(d), the weight of incorrectly classified events, for which $y_n f_k(x_n) = -1$, is *increased*, while that of correctly classified events, for which $y_n f_k(x_n) = +1$, is *decreased*.

`AdaBoost` is a rather cryptic algorithm, which, like decision tree classifiers, does not seem to fit into the general discussion about average loss given above. However, subsequent to the publication of the `AdaBoost` algorithm, Friedman, Hastie, and Tibshirani [30] showed that this algorithm can be viewed as a way to minimize the average loss function

$$E[f] = \int dx \int dy \exp[-yf(x)]\, p(y,x), \tag{66}$$

whose minimum occurs at

$$f(x) = \frac{1}{2} \ln \frac{p(y=+1|x)}{p(y=-1|x)}. \tag{67}$$

To see this consider the problem of minimizing the error rate on the training sample,

$$\epsilon = \frac{1}{N} \sum_{n=1}^{N} I[y_n f(x_n)].$$

Minimizing the error rate directly in a reasonable amount of time is extremely difficult, therefore, in practice, a proxy for the error rate is minimized instead. Noting that $\exp(-y_n f(x_n)) > 1$ when $y_n f((x_n) < 0$, one such proxy is the righthand side of the following

$$\epsilon = \frac{1}{N} \sum_{n=1}^{N} I[y_n f(x_n)],$$

$$\leq \frac{1}{N}\sum_{n=1}^{N} e^{-y_n f(x_n)} I[y_n f(x_n)],$$
$$\leq \frac{1}{N}\sum_{n=1}^{N} e^{-y_n f(x_n)}. \tag{68}$$

Note that in the limit $N \to \infty$, the righthand side of the above equation, which is an upper bound on the error rate, converges to Eq. (66). Furthermore, from the recursive definition of the normalized event weights $w_{k+1,n} = w_{k,n}\exp(-\alpha_k y_n f_k(x_n))/Z_k$ in the `AdaBoost` algorithm, we conclude that

$$\epsilon \leq \frac{1}{N}\sum_{n=1}^{N} e^{-y_n f(x_n)} = \prod_{k=1}^{K} Z_k,$$
$$= \prod_{k=1}^{K}\sum_{n=1}^{N} \omega_{k,n}\exp(-\alpha_k y_n f_k(x_n)) \equiv \epsilon'. \tag{69}$$

If we regard the coefficients α_k as free parameters (by neglecting the dependence of the event weights on α_k), we can minimize ϵ' with respect to α_k by solving

$$\frac{\partial \epsilon'}{\partial \alpha_k} = -\left(\prod_{j\neq k} Z_j\right)\sum_{n=1}^{N} \omega_{k,n} y_n f_k(x_n)\exp(-\alpha_k y_n f_k(x_n)) = 0,$$
$$\text{that is, } \sum_{n=1}^{N} \omega_{k,n} y_n f_k(x_n)\exp(-\alpha_k y_n f_k(x_n)) = 0. \tag{70}$$

Since $y \in \{-1, +1\}$, we can write

$$e^{-\alpha_k}\sum_{n=1}^{N} \omega_{k,n} I[y_n f_k(x_n)] - e^{\alpha_k}\sum_{n=1}^{N} \omega_{k,n} I[-y_n f_k(x_n)] = 0,$$
$$e^{-\alpha_k}(1-\epsilon_k) - e^{\alpha_k}\epsilon_k = 0,$$
$$\text{where, recall, } \epsilon_k = \sum_{n=1}^{N} \omega_{k,n} I[-y_n f_k(x_n)], \tag{71}$$

is the weighted error rate. We therefore conclude that the upper bound on the error rate ϵ' is minimized if we choose the coefficients to be $\alpha_k = \frac{1}{2}\ln[(1-\epsilon_k)/\epsilon_k]$, which is indeed the choice made in `AdaBoost`.

Therefore, in spite of appearances, boosted decision trees fit into the mathematical framework sketched above. In particular, `AdaBoost` can be viewed as a clever way to minimize the average exponential loss given in Eq. (66). Moreover, while the boosted decision tree $f(x)$ cannot be interpreted as a probability, it can be mapped to a probability by inverting Eq. (67),

$$p(y = +1|x) = \frac{1}{1+\exp(-2f(x))}. \tag{72}$$

Below, we illustrate the use of the `AdaBoost` algorithm using the Toolkit for Multivariate Analysis `TMVA` [31], which is released with the `ROOT` [32] package from CERN. Note, in the `TMVA` implementation, α_k is defined omitting the factor of 1/2, therefore, in order to convert the unnormalized BDT, $f(x)$, in `TMVA` to a probability, the appropriate mapping is

$$p(y = +1|x) = \frac{1}{1+\exp(-f(x))} \quad (\texttt{TMVA}). \tag{73}$$

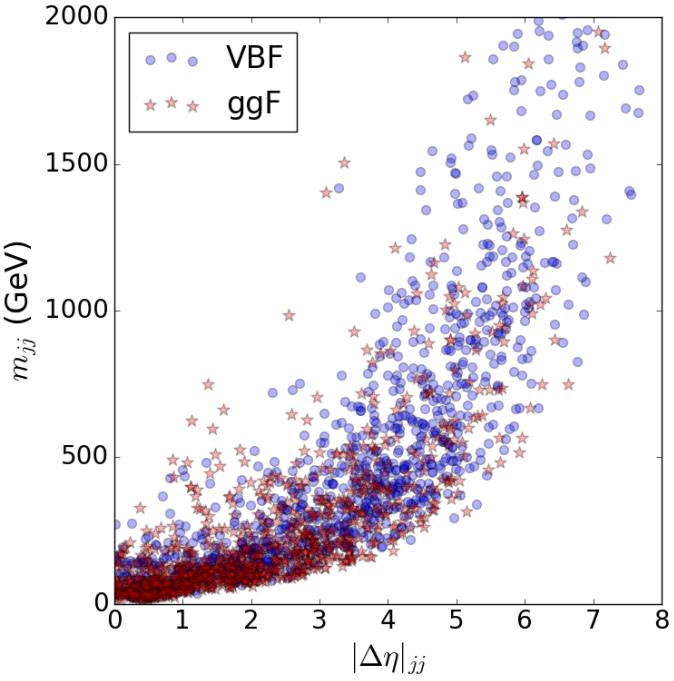

Fig. 12: Simulated distributions of the discriminating variables $(|\Delta\eta_{jj}|, m_{jj})$ for VBF and ggF events. As expected, there is a larger rapidity gap between the jets inVBF events than those in ggF, which arise from gluon radiation.

VBF/ggF discrimination

In this example, a BDT is trained using the `AdaBoost` algorithm in `TMVA` to discriminate between events in which the Higgs boson is created via vector boson fusion (VBF) and events in which the Higgs boson is created via gluon gluon fusion (ggF). The key difference between VBF events and ggF events is that the former features a pair of forward (i.e., large rapidity) jets that is absent from the latter. It is found that the two most discriminating variables between these two classes of events are the absolute pseudo-rapidity difference $|\Delta\eta|_{jj}$ between the two jets and the associated di-jet mass m_{jj} and. The predicted distributions of the two variables is shown in Fig. 12.

We use a training sample size of $N = 20,000$ events, split equally between VBF and ggF events with assigned targets of $y = +1$ and $y = -1$, respectively. The `TMVA` training parameters are `BoostType=AdaBoost`, `NTrees=800`—the number of trees K, `nEventsMin=100`—the minimum number of events per bin, and `nCuts=50`—the number of binary partitions per variable to search for the optimal partition, i.e., cut. The optimal cut is the one which gives the greatest *decrease* in impurity as measured by the Gini index[7], defined by $p(1-p)$ where $p = S/(S+B)$ is the purity and S and B are the signal and background counts, respectively, in a given bin. A bin is maximally pure, either pure signal or pure background, when the Gini index is zero.

Figure 13 shows the first six decision trees as histograms, each with its associated coefficient $\alpha_k = \ln[(1-\epsilon_k)/\epsilon_k]$[8] printed on the histogram. A decision tree is a piecewise constant function in which each bin (i.e., leaf) is assigned a value. In the `AdaBoost` algorithm, the values are $y = \pm 1$; in our example, $y = -1$ for bins in which $B > S$ (i.e., ggF bins) and $+1$ for bins in which $S > B$ (i.e., VBF bins). A given feature vector $x = |\Delta\eta|_{jj}, m_{jj}$, characterizing an event, will fall in a bin in each of the six decision trees of Fig. (13) and the BDT is equal to the average $\sum_{k=1}^{6} f_k(x)$ where each tree $f_k(x)$

[7] After Italian statistician Corrado Gini, 1884-1965.

[8] As noted, `TMVA` omits the factor of 1/2.

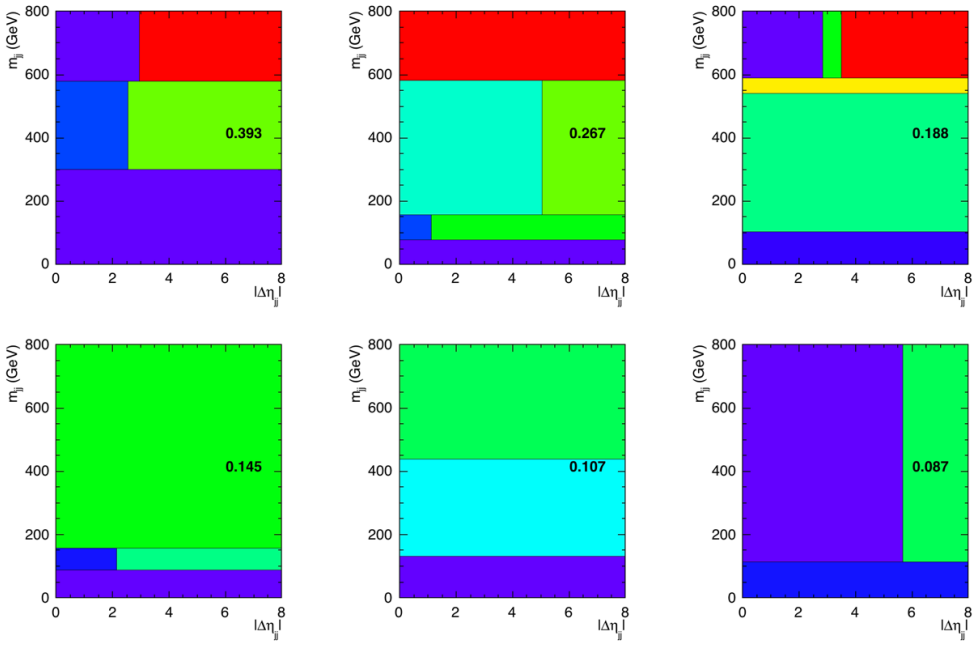

Fig. 13: The first six of the 800 decision trees, displayed as 2D histograms, showing the coefficients $\alpha_1, \cdots, \alpha_6$ associated with the threes.

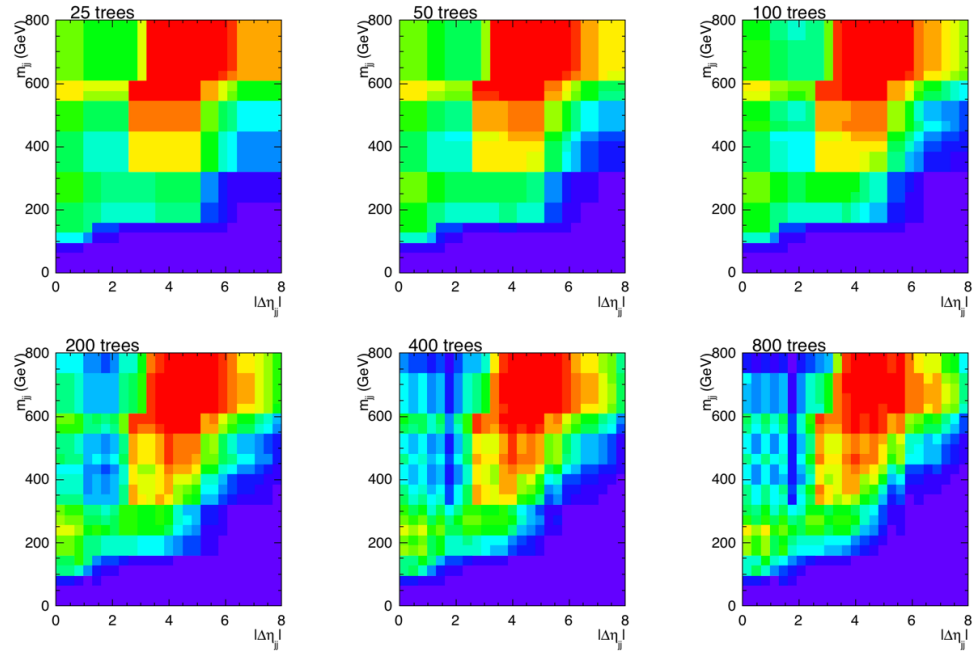

Fig. 14: The outputs of boosted decision trees averaged over differing numbers of decision trees, 25, 50,...,800. Each $BDT(x)$, with $x = |\Delta\eta|_{jj}, m_{jj}$, is mapped to the probability $p(y = +1 \,|\, x) = 1/[1 + \exp(-BDT(x))]$.

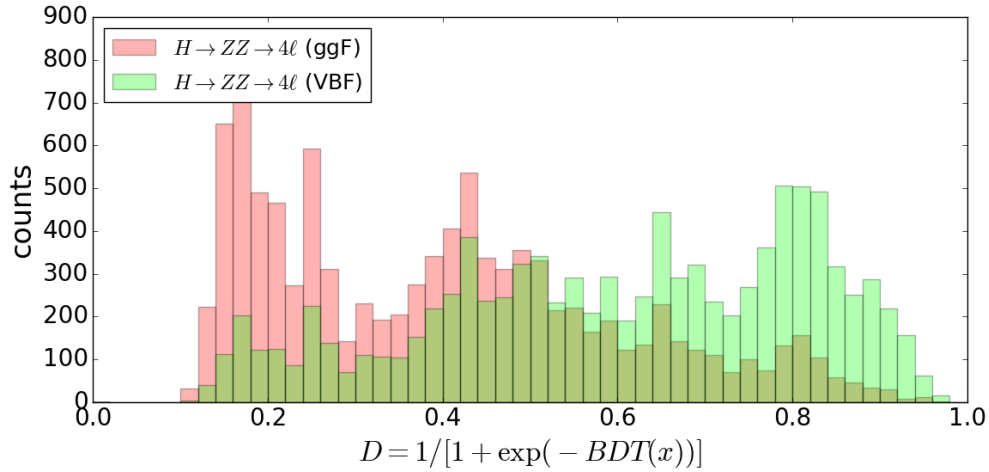

Fig. 15: The distributions of the discriminant $D(x) = 1/[1 + \exp(-BDT(x))]$, where $BDT(x)$ is a boosted decision tree with $K = 800$ trees.

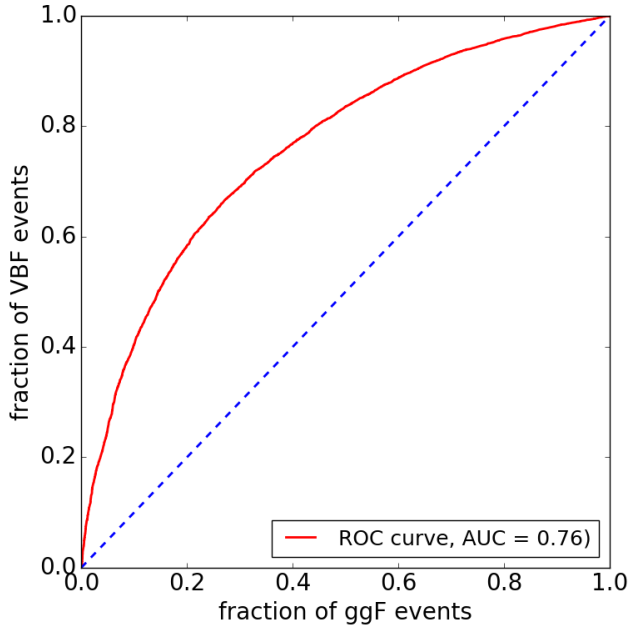

Fig. 16: Receiver operating characteristic (ROC) curve. The area under the curve (AUC) is a commonly used global measure of the discrimination power of a classifier.

returns either $+1$ or -1 depending on the bin in which x falls. In other words, a BDT is an average over histograms, each with different set of bins. While the piecewise constant nature remains, the more histograms (that is, trees) are averaged, the smoother one expects the BDT output to become. This is illustrated in Fig. 14, which shows the effect of averaging over an increasing number of trees. Finally, Figs. 15 and 16 show the distribution of the BDT, in which the output has been mapped to the probability $p(\text{VBF}\,|\,x) \equiv p(y = +1\,|\,x) = 1/[1+\exp(-BDT(x))]$, and the receiver operating characteristic (ROC) curve of the BDT.

The ROC curve, and the area underneath it (AUC), are often used as simple measures of the performance of a binary classifier. The larger the AUC the better the performance of the classifier.

Summary

We have given an overview of the frequentist and Bayesian approaches to statistical inference and a brief survey of the main mathematical ideas that underpin supervised machine learning. Frequentist analysis is based on the relative frequency interpretation of probability and, ideally, adheres to the frequentist principle: repeated application of a statistical procedure will yield statements a fraction $f \geq p$ of which are guaranteed to be true, where p is the desired confidence level. The Bayesian approach uses the degree of belief interpretation of probability and Bayes theorem as the primary inference algorithm. In both approaches, the key task is building an accurate probability model.

A brief introduction to supervised machine learning was given in which the emphasis was clarifying the critical role of the loss function. We noted the mathematical fact that the quantity approximated by a machine learning model is determined by the loss function and not by the particulars of the model provided that sufficient training data are used, the model is sufficiently flexible, and a good approximation to the minimum of the average loss can be found.

Acknowledgement

I thank Nick Ellis, Martijn Mulders, Kate Ross, and their counterparts from JINR, for organizing and hosting a very enjoyable school, and the students for their keen participation and youthful enthusiasm. These lectures were supported in part by US Department of Energy grant DE-SC0010102.

References

[1] L. Lyons, *Statistics for Nuclear and Particle Physicists*, Cambridge University Press, Cambridge (1989).

[2] F. James, *Statistical Methods in Experimental Physics*, 2nd Edition, World Scientific, Singapore (2006).

[3] G. Cowan, *Statistical Data Analysis*, Oxford University Press, Oxford (1998).

[4] R. J. Barlow, *Statistics: A Guide To The Use Of Statistical Methods In The Physical Sciences*, The Manchester Physics Series, John Wiley and Sons, New York (1989).

[5] S. K. Chatterjee, *Statistical Thought: A Perspective and History*, Oxford University Press, Oxford (2003).

[6] L. Daston, "How Probability Came To Be Objective And Subjective," Hist. Math. 21, 330 (1994).

[7] S. Chatrchyan *et al.* [CMS Collaboration], "Measurement of the properties of a Higgs boson in the four-lepton final state," Phys. Rev. D **89**, no. 9, 092007 (2014) doi:10.1103/PhysRevD.89.092007 [arXiv:1312.5353 [hep-ex]].

[8] J. Neyman, "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," Phil. Trans. R. Soc. London A236, 333 (1937).

[9] G. J. Feldman and R. D. Cousins, "Unified approach to the classical statistical analysis of small signals," Phys. Rev. D57, 3873 (1998).

[10] S. E. Fienberg and D. V. Hinkley, eds., *R.A. Fisher: An Appreciation*, Lecture Notes on Statistics, Volume 1, Springer Verlag (1990).

[11] G. Cowan, K. Cranmer, E. Gross, O. Vitells "Asymptotic formulae for likelihood-based tests of new physics," Eur. Phys. J. C71, 1554 (2011).

[12] G. Taraldsen and B.H. Lindqvist, "Improper Priors Are Not Improper," The American Statistician, Vol. 64, Issue 2, 154 (2010).

[13] G. Aad *et al.* [ATLAS Collaboration], "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," Phys. Lett. B **716**, 1 (2012) [arXiv:1207.7214 [hep-ex]].

[14] S. Chatrchyan *et al.* [CMS Collaboration], "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," Phys. Lett. B **716**, 30 (2012) [arXiv:1207.7235 [hep-ex]].

[15] H. Jeffreys, *Theory of Probability*, 3rd Edition, Clarendon Press, Oxford (1961).

[16] V. M. Abazov *et al.* [D0 Collaboration], "Observation of Single Top Quark Production," Phys. Rev. Lett. **103**, 092001 (2009) [arXiv:0903.0850 [hep-ex]].

[17] T. Aaltonen *et al.* [CDF Collaboration], "First Observation of Electroweak Single Top Quark Production," Phys. Rev. Lett. **103**, 092002 (2009) [arXiv:0903.0885 [hep-ex]].

[18] S. Sekmen *et al.*, "Interpreting LHC SUSY searches in the phenomenological MSSM," JHEP **1202**, 075 (2012) doi:10.1007/JHEP02(2012)075 [arXiv:1109.5119 [hep-ph]].

[19] V. Khachatryan *et al.* [CMS Collaboration], "Phenomenological MSSM interpretation of CMS searches in pp collisions at sqrt(s) = 7 and 8 TeV," JHEP **1610**, 129 (2016) doi:10.1007/JHEP10(2016)129 [arXiv:1606.03577 [hep-ex]].

[20] V. Khachatryan *et al.* [CMS Collaboration], "Search for supersymmetry in pp collisions at sqrt(s) = 8 TeV in final states with boosted W bosons and b jets using razor variables," Phys. Rev. D **93**, no. 9, 092009 (2016) doi:10.1103/PhysRevD.93.092009 [arXiv:1602.02917 [hep-ex]].

[21] L. Demortier, S. Jain and H. B. Prosper, "Reference priors for high energy physics," Phys. Rev. D **82**, 034002 (2010) [arXiv:1002.1111 [stat.AP]].

[22] I. J. Myung, V. Balasubramanian, and M. A. Pitt, "Counting probability distributions: Differential geometry and model selection," PNAS, **97** 11170-11175 (2000); doi: 10.1073/pnas.170283897.

[23] A. Turing, "Computing Machinery and Intelligence," Mind **59** 433-460 (1950).

[24] D. Silver et al., "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," Science **362** 1140-1144 (2018); DOI: 10.1126/science.aar6404.

[25] Ruck et al., IEEE Trans. Neural Networks 4, 296-298 (1990).

[26] Wan, IEEE Trans. Neural Networks 4, 303-305 (1990).

[27] Richard and Lippmann, Neural Computation. 3, 461-483 (1991).

[28] H. J. Yang, B. P. Roe and J. Zhu, "Studies of boosted decision trees for MiniBooNE particle identification," Nucl. Instrum. Meth. A **555**, 370 (2005) doi:10.1016/j.nima.2005.09.022 [physics/0508045].

[29] Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and Sys. Sci. **55** (1), 119 (1997).

[30] J. Friedman, T. Hastie and R. Tibshirani, ï£¡Additive logistic regression: a statistical view of boosting,ï£¡ The Annals of Statistics, **28** (2), 377-386 (2000).

[31] P. Speckmayer, A. Hocker, J. Stelzer and H. Voss, "The toolkit for multivariate data analysis, TMVA 4," J. Phys. Conf. Ser. **219**, 032057 (2010). doi:10.1088/1742-6596/219/3/032057

[32] Rene Brun and Fons Rademakers, "ROOT - An Object Oriented Data Analysis Framework," Proceedings AIHENP 96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A **389**, 81-86 (1997). See also https://root.cern.ch/.