Section 1

Profile likelihood method

In this appendix, we briefly review the the profile likelihood method used in Sec. **??**. In particular, we describe the motivation and justification to consider this method.

First of all, the experimental outcome of our concern can be expressed as a set of random variables $\boldsymbol{x} \equiv \{x_1, \cdots, x_n\}$, with n being the number of observables. The distribution of these variables is due to both the intrinsic physical randomness (*i.e.*, the statistical fluctuation) and the uncertainty in detector responses such as the efficiency, momentum reconstruction, and so on. We assume $\boldsymbol{x}$ obey some probability distribution function and express it as $f(\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_m\}$ parametrize (in many cases unknown) uncertainties listed above. When we repeat N experiments and obtained N sets of observables expressed as $\boldsymbol{x}^a$ $(a = 1, \cdots, N)$, we define the likelihood function L as

$$L = \prod_{a=1}^{N} f(\boldsymbol{x}^a; \boldsymbol{\theta}). \tag{1.1}$$

Since L should take a relatively larger value if the assumed distribution f approximates the reality very well, we may perform the maximization of L against the choice of $\boldsymbol{\theta}$ to obtain the correct probability distribution. Such maximization procedure can be performed analytically only for several simple distribution functions. Thus, in many cases, we need a numerical calculation of the maximization procedure, which can be performed with the `MINUIT` package [**?**].

In our analysis in Sec. **??**, the data is given in the form of the histogram. In this case, x_i $(i = 1, \cdots, n)$ denotes the observed number of events in each bin labeled by i, with n being the number of bins. The product of the probability distribution function for each bin gives f, which is equivalent to L in this case, expressed as

$$L(\boldsymbol{x}; \boldsymbol{\theta}) \equiv f(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\left[-\frac{(x_i - \mu_i(\boldsymbol{\theta}))^2}{2x_i}\right], \tag{1.2}$$

with $\mu_i(\boldsymbol{\theta})$ being the average number of events of the bin i calculated using the parameters $\boldsymbol{\theta}$. Note that $x_i \gg 1$ is assumed for each bin and the central limit theorem is used to replace the Poisson to the Gaussian distribution. Then, it is clear that the maximization of L is equivalent to the minimization of χ^2 defined as

$$\chi^2 \equiv -2 \ln L(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(x_i - \mu_i(\boldsymbol{\theta}))^2}{x_i}, \tag{1.3}$$

which is the so-called Neyman's χ^2 variable. χ^2 obeys the chi-square distribution when the distribution of x_i is well approximated by the Gaussian, and can be easily used to find the optimized choice of parameters $\boldsymbol{\theta}$ and their errors.

Similarly, one can apply the likelihood maximization to the test of the model. Let $\boldsymbol{\theta}_{\text{true}}$ and $\boldsymbol{\theta}_{\text{test}}$ be the model in the reality and that we want to test, respectively. For example, in the new physics search, the former corresponds to the new physics model, while the latter to the SM. Then we can define the test-statistic

$$q(\boldsymbol{\theta}_{\text{test}}) = -2\ln \frac{L(\boldsymbol{x}; \boldsymbol{\theta}_{\text{test}})}{L(\boldsymbol{x}; \boldsymbol{\theta}_{\text{true}})}, \tag{1.4}$$

which plays a role of the so-called $\Delta\chi^2$ variable according to the discussion above. Again q may obey a chi-square distribution with some degrees of freedom, and can be used to obtain sensitivities to the new physics, *i.e.,* the $95\,\%$ C.L. exclusion and the 5σ discovery. Note that the denominator of the test static can also be expressed as $L(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$, where the hat denotes the values of $\boldsymbol{\theta}$ that maximize the function L.

However, the situation may be more complicated since some of the parameters $\boldsymbol{\theta}$ are not directly related to the model parameters, but express the background yield, detector effects, systematic errors, and so on, which should be determined from experimental data. Such additional parameters are often called nuisance parameters. To treat nuisance parameters, it is convenient to rely on the profile likelihood method [?].

For this purpose, we divide the parameters into two categories: the model parameter of our interest $\boldsymbol{\mu}$ and nuisance parameters $\boldsymbol{\theta}$. Similarly to the discussion without the nuisance parameters, let $\boldsymbol{\mu}$ be the model in we want to test. The test static is defined as

$$q(\boldsymbol{\mu}) = -2\ln \frac{L(\boldsymbol{x}; \boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}))}{L(\boldsymbol{x}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})}, \tag{1.5}$$

where the meaning of the hat is the same as above, while $\hat{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu})$ denotes the values that maximize L with fixed values of $\boldsymbol{\mu}$. The motivation for this choice is given by the Wilk's theorem [?], which proves that $q(\boldsymbol{\mu})$ asymptotically obeys the chi-square distribution whose degrees of freedom equal to the number of model parameters $\boldsymbol{\mu}$. Note that this the statement is highly non-trivial, since the individual term $-2\ln L$ *does not* obey chi-square distributions in this case. Thanks to the theorem, we can perform the same analysis under the existence of nuisance parameters and, in particular, absorb the effects of systematic errors into the choice of parameters $\boldsymbol{\theta}$.

(♣ **Comment on distribution of nuisance parameters?** ♣)