# Quasi-optimal weights: a versatile tool of data analysis

**Fyodor Tkachov**

Institute for Nuclear Research, Russian Academy of Sciences, Moscow 117312
Russia

E-mail: ftkachov@inr.ac.ru

**Abstract.** In this first part of the two-part account of the enabling technologies behind a successful completion of the Troitsk-nu-mass experiment, the parameter estimation method of quasioptimal weights is reviewed. In regard of statistical quality, it is on a par with the maximal likelihood method. but exceeds the latter in analytical transparency, flexibility, scope of applicability and numerical robustness. It also couples perfectly with the optimal jet definition and thus provides a comprehensive framework for data processing in particle physics and beyond.

## 1. Introduction

The currently best ν mass bound was published by the Troitsk-ν-mass experiment [1]. This happened after the experiment had been plagued by the so-called “Troitsk anomaly” for about 10 years. The anomaly went away after a reanalysis that was made possible by two enabling technologies: the statistical method of quasioptimal weights for parameter estimation [2] and the Oberon technologies for software development. The latter are discussed in a companion talk [3]. The present talk reviews the former.

It should be noted that the two technologies per se did not solve any specific physics problem, but provided a flexible and efficient framework to experiment with and implement specific improvements of the analysis in a real-life context under real-life limitations. Everything could in theory have been done with maximal likelihood and fortran or even C++, but at a cost that proved prohibitive in that context. A proper *technique* is about minimizing effort while attaining the best result, it is rooted in how well one understands the basic principles of what one works with.

## 2. The method

A basic problem of mathematical statistics is, given a parameterized distribution $\pi_M(\mathbf{P})$ and a random sample $\{\mathbf{P}_i\}_{i=1\dots N}$, to estimate the unknown value $M_0$ of the parameter $M$. One seeks to obtain an estimate $M_*$ for the parameter and a standard error estimate $\sigma_*$ (or the corresponding confidence intervals).

The popular least squares method does not provide a full control; in fact, this project was prompted in the Fall of 2005 by a doubt expressed by the leader of the Troitsk-ν-mass experiment, the late Vladimir Lobashev as to whether the Poisson-distributed background was treated correctly by least squares.

The simple and transparent method of moments is universally regarded as inferior and mostly neglected in data processing applications. Among the few exceptions are [4] and [5]. Both turn out to be special cases of the method of quasi-optimal weights.

The maximal likelihood method, if applicable, yields the best possible estimate, obtained by maximizing the likelihood function,

$$\sum_i \ln \pi_M(\mathbf{P}_i) \le \sum_i \ln \pi_{M_{\rm FFRC}}(\mathbf{P}_i). \tag{1}$$

The estimate corresponds to the fundamental Fisher-Frechet-Rao-Cramer bound:

$$\sigma^2_{\rm FFRC} = \frac{1}{N}\left[\text{Fisher's information}\right]^{-1} \tag{2}$$

Note that the maximum can be found by solving

$$\frac{\partial}{\partial M}\sum_i \ln \pi_M(\mathbf{P}_i) = 0 \tag{3}$$

But what if $\pi_M(\mathbf{P})$ is unknown as a formula? This is often the case in high energy physics where only a Monte Carlo event generator may be available but not an explicit expression for the probability distribution (this is due to a very high dimensionality of the event space). On top of that, HEP data processing involves a heavy dose of chiropractic (event selection, jet algorithms, etc.).

Consider the method of generalized moments. For any weight ("generalized moment") $f(\mathbf{P})$, its mean value is a function of the unknown parameter

$$\langle f\rangle = \int d\mathbf{P}\, f(\mathbf{P})\,\pi_M(\mathbf{P}) = F(M) \tag{4}$$

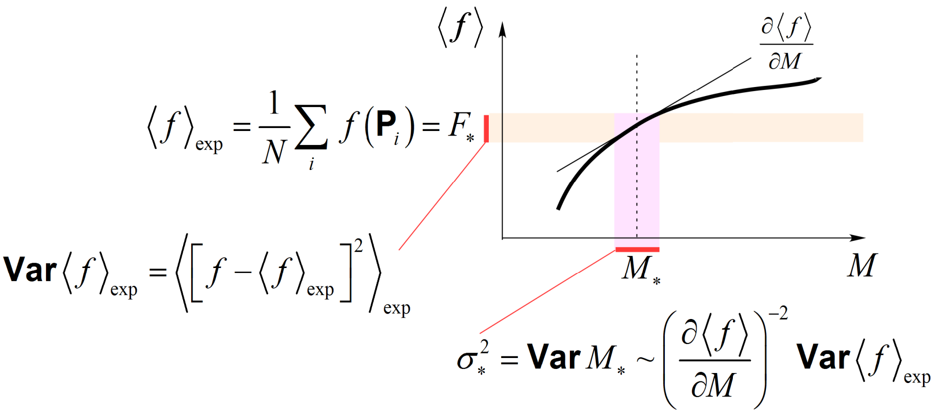
On the experimental side, one can estimate this mean as follows

$$\langle f\rangle_{\rm exp} = \frac{1}{N}\sum_i f(\mathbf{P}_i) = F_* \tag{5}$$

One then estimates the parameter,

$$M_* = F^{-1}(F_*) \tag{6}$$

This is easily visualized:


$\langle f\rangle$
$\frac{\partial\langle f\rangle}{\partial M}$
$\langle f\rangle_{\rm exp} = \frac{1}{N}\sum_i f(\mathbf{P}_i) = F_*$
$\mathbf{Var}\langle f\rangle_{\rm exp} = \left\langle\left[f-\langle f\rangle_{\rm exp}\right]^2\right\rangle_{\rm exp}$
$M_*$
$M$
$\sigma_*^2 = \mathbf{Var}\,M_* \sim \left(\frac{\partial\langle f\rangle}{\partial M}\right)^{-2}\mathbf{Var}\langle f\rangle_{\rm exp}$

Note that both error estimates are easily evaluated. Note also that

$$\mathbf{Var}\langle f\rangle_{\rm exp} \sim \frac{1}{N}\mathbf{Var}\langle f\rangle, \quad \mathbf{Var}\langle f\rangle = \left\langle\left[f-\langle f\rangle\right]^2\right\rangle \tag{7}$$

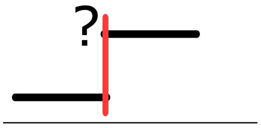
Originally Pearson (1894) dealt with a one-dimensional event space and used:

$$f(\mathbf{P}) = \mathbf{P}^n \tag{8}$$

which yielded suboptimal estimates compared with ML.

However, the method is analytically transparent unlike either of the alternatives, and there are deep mathematical reasons to consider weights. Indeed, an alternative to the set-theoretic notion of a general function is one based on weighted values rather than point values of the function. In this alternative scheme the basic notions are choosen to be functions/mappings (with compositions, morphisms, categories etc.) rather than (sub)sets (with the corresponding set operations such as intersections etc.). An accessible review of these matters is given in [6]. It should be emphasized that such a functional view of the mathematical analysis is as comprehensive as the set-theoretic one. But the heuristics of the two schemes are entirely different: the weighted values are an artificial device in the set-theoretic view whereas in the functional view they are a natural starting point.

In the conventional view, a function is a correspondence $x \to g(x)$. This, however, is fully meaningful only for continuous $g(x)$. There is no natural way to define the value at a discontinuity.

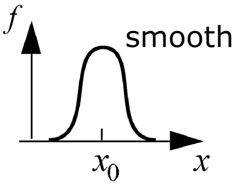


(9)

In that regard a more satisfactory way is to define a function via its weighted values

$$f \to \langle g\, f \rangle \quad \left[\!\left[ = \int \mathrm{d}x\, g(x)\, f(x) \right]\!\right] \tag{10}$$

where *f* are the so-called test functions usually chosen to be smooth (or at least continuous):



(11)

Such a definition allows one to recover the function values at the points of continuity but the problem of values at discontinuities is bypassed altogether.

In particular — and immediately relevant to approximation methods — instead of comparing point values, one now compares weighted/smeared values. This directly leads to the powerful Bubnov-Galerkin method very widely used in engineering and other numerical applications (see e.g. [7]).

The functional viewpoint singles out the method of generalized weights as a fundamental starting point of any thinking about parameter estimation. So, let us look at it more closely, putting aside the mantra of its inferiority.

One has a simple and explicit expression for the error estimate:

$$\sigma_*^2 \sim \frac{1}{N} \left\{ \left( \frac{\partial \langle f \rangle}{\partial M} \right)^{-2} \left\langle \left[ f - \langle f \rangle \right]^2 \right\rangle \right\} \tag{12}$$

Let us minimize it. From the minimum equation

$$\frac{\delta}{\delta f(\mathbf{P})} \{\dots\} = 0 \tag{13}$$

one easily finds

$$f_{\text{opt}}(\mathbf{P}) = \frac{\partial}{\partial M} \ln \pi_M(\mathbf{P}) \tag{14}$$

This point turns out to be a true minimum, with Fisher's information as the minimum value. This exercise was performed in [8].

However, the optimal weight depends on *M*, and its actual value $M_0$ is unknown. Then an iterative procedure naturally suggests itself:

$$M_i \rightarrow f_{\text{opt},M_i}(\mathbf{P}) \rightarrow M_{i+1} \tag{15}$$

The problem of choosing the initial value $M_1$ is not specific to this method; it is implicit in any method based on optimizations such as ML and least squares.

Note that each $M_i$ is a correct statistical estimate per se, only the corresponding errors may be suboptimal. Note also that the convergence of the sequence $M_i$ is not an issue here since it is the variance that the iterative procedure is intended to minimize.

The non-optimal weights $f_{\text{opt},M_i}(\mathbf{P})$ are close to the optimal one according to how $M_i$ is close to $M_0$, whence the name of the method.

Given that the minimum of the variance in the space of weights is quadratic, there is a freedom in the choice of quasi-optimal weights, and the method is rather less demanding in regard of the knowledge of the underlying probability density than the ML method.

That the iterative procedure is equivalent to ML is seen from the fact that

$$\langle f_{\text{opt}} \rangle = 0\,. \tag{16}$$

(This assumes that the optimal weight and the averaging are taken at the same *M*.)

The iterative procedure is similar to the optimizations involved in both the ML and least squares methods. It appears that the method of quasi-optimal weights reveals and efficiently exploits an analytical structure that have been there all along but remained eclipsed by the glory of the maximal likelihood. (One might also say that it fell victim to the personal rivalry between Sir Robert Fisher and Karl Pearson.)

**Example 1**

Some special cases of the quasi-optimal weighting have been known for a while. The case of linear dependence of the probability distribution on *M* has been known since 1992 [4]

$$\pi(\mathbf{P}) = \pi_0(\mathbf{P}) + g\pi_1(\mathbf{P}), \quad f_{\text{opt}}(\mathbf{P}) = \frac{\pi_1(\mathbf{P})}{\pi(\mathbf{P})} \tag{17}$$

This has been used in a number of weak signal searches.

This is closely related to the following criterion for hypothesis testing which is locally most powerful near $g = 0$: $\pi_1(\mathbf{P}) / \pi_0(\mathbf{P})$.

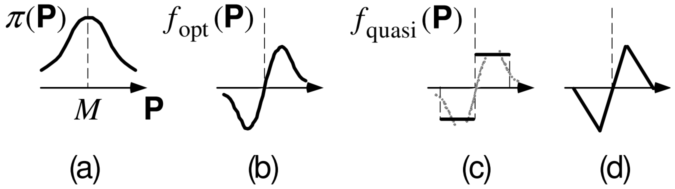
Another special case was constructed by Jorg Pretz within the context of the g-2 experiments [5].

**Example 2**

Consider the Cauchy/Breit-Wigner distribution. It has no mean yet the quasi-optimal weights work just fine to determine *M*:

$$\pi(\mathbf{P}) \propto \frac{1}{(M-\mathbf{P})^2 + \Gamma^2}, \quad f_{M,\text{opt}}(\mathbf{P}) \propto \frac{M-P}{(M-\mathbf{P})^2 + \Gamma^2} \tag{18}$$

In this simple example one can see the various options for an approximate choice of the quasi-optimal weight which, it should be emphasized, need not be chosen to exactly coincide with the analytical formula:



(19)

Any of these shapes would yield a valid estimate for *M*.

**Example 3**
Compare Poisson and normal distributions. This concerns situations where one deals with a Poisson distributions but implicitly assumes normality of all distributions in the actual data processing.

Consider the normal distribution

$$\pi_\mu(\mathbf{P}) \propto \exp\left[-(P-\mu)^2/2\sigma^2\right] \tag{20}$$

The two optimal weights for the two parameters are as follows:

$$f_{\mathrm{opt},\,\mu} \propto (P-\mu), \quad f_{\mathrm{opt},\,\sigma} \propto (P-\mu)^2 \tag{21}$$

Now consider the Poisson distribution with an integer *n* in place of the real *P*:

$$\pi_\mu(n) = e^{-\mu}\frac{\mu^n}{n!}, \quad f_{\mathrm{opt},\,\mu}(n) = \left(\frac{n}{\mu}-1\right) \tag{22}$$

Using methods devised for the normal distribution to estimate the Poisson parameter $\mu$ is similar to working with the distribution

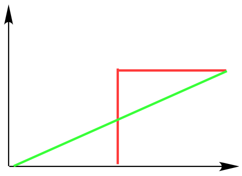
$$\pi_\mu(n) \propto \exp\left[-(n-\mu)^2/2\mu^2\right] \tag{23}$$

and the corresponding family of weights is

$$f_{\mathrm{opt},\,\mu}(n) \propto \left(\frac{n}{\mu}-1\right) + \frac{1}{2}\left(\frac{n}{\mu}-1\right)^2 \tag{24}$$

The difference from the purely Poisson weight is indicative of a bias introduced thereby. Numerically, for $\mu \sim 0.15$–1.5, the effect is 10–15%, which is large enough to be of concern.

**Example 4**
Consider how the continuous weights compare with event selection cuts widely used in HEP experiments. Consider a one-dimensional event space [0,1] and the linear probability density:



(25)

Consider the variances of the linear weight and the shown cut relative to the linear probability distribution. The ratio of the two variances is 3, and the ratio of the corresponding sigmas is 1.71. In other words, replacing the cut with a continuous weight in this case converts a $3\sigma$ signal into a $5\sigma$ one.

**Example 5**
Our last example is the observation of the top quark in all-hadrons channel by D0 as described in [10]. A continuous weight (“an average jet count’) was devised empirically and proved to be “particularly powerful” in a multivariate analysis with thirteen variables.

Even if a formula such as $f_{\mathrm{opt}}(\mathbf{P}) = \frac{\partial}{\partial M}\ln \pi_M(\mathbf{P})$ remains somewhat of a theoretical fiction in the case of multijet event space with a Monte Carlo event generator, an explicit formula is always a great force as it sheds light on what otherwise is a trial-and-error search.

## 3. The case of non-uniform event sample

The method can be extended to the case of non-uniform samples [2].

A set of values $U_i$ of a control parameter is chosen, and for each value a measurement $P_i$ is taken distributed according to a probability distribution $\pi_{i,\theta}(P)$. All these are assumed to depend on the unknown parameter to be estimated accroding to a known formula. If all $P_i$ for different $U_i$ are independent then the situation reduces to the simple case by defining a composite event $\otimes_i P_i$ and the corresponding probability density $\otimes_i \pi_{i,\theta}(P_i)$. The optimal weight then is:

$$\varphi_{\text{opt}}(\{P_i\}) = \sum_i \frac{\partial}{\partial\theta} \ln \pi_{i,\theta}(P_i) = \sum_i \varphi_{\text{opt}_i}(P_i) \tag{26}$$

with the explicit formula for variance:

$$\textbf{Var}\,\varphi_{\text{opt}} = \sum_i \left[ \varphi_{\text{opt}_i}(P_i) - \left\langle \varphi_{\text{opt}_i} \right\rangle_{\theta_{\text{exp}}}^{\text{th}} \right]^2 \tag{27}$$

All this is straightforward to code.

## 4. Notes on usage in the Troitsk-ν-mass experiment

A number of Monte Carlo tests showed that the method is indeed asymptotically equivalent to ML (which fact is quite obvious from the analytical formulae of the method).

Programming implementation of the method is quite straightforward, although some flexibility of the implementation language is desirable (actually used was a minimalistic strictly type-safe language Oberon/Component Pascal, a direct descendant of Pascal, see the companion talk for details [3]).

Numerically, one has to solve a system of equations, which is a simpler task than an optimum search (for a fast Newton-type algorithm, one only needs first derivatives of the mapping involved in the former case, whereas the optimum search requires second derivatives).

The method is less sensitive than, say, MINUIT to the “narrow valleys” situation (although this difficulty cannot be eliminated altogether), as well as to non-smoothness of the functions involved.

One can easily accomodate knowledge about the nature of the probability distributions involved (Poisson etc.).

With a poor statistics, the resulting multidimensional systems of equations may not have a solution; in practice this was not a problem (mismatches that were encountered were small more than enough not to affect the results) but some (theoretical) research may be called for. Similarly, a Student-type correction factors for confidence intervals may be required in the poor statistics case (this problem, however, is a general one and not limited to the quasi-optimal weights; in this regard the present method is neither better nor worse than everything else).

The same conceptual framework earlier yielded the so-called Optimal Jet Definition [11] which was the first to demonstrate an *O*(*N*) behavior with respect to the number of particles in the final state [12]. The OJD remains the only jet definition derived from first principles in a systematical scientific fashion, without ad hoc constructions. The fact that it meshes well with quasi-optimal weights bodes well for the complicated experimental situations such as the abovementioned one encountered by D0 in the all-jets channel.

Lastly, a fundamental nature of the argument behind OJD is underscored by the fact that a very similar derivation yielded an efficient multi-dimensional density modelling algorithm [13].

To summarize, a general and flexible method of quasi-optimal weights (together with a closely related optimal jet definition, the general multi-dimensional density modelling algorithm, etc.) offers a rather comprehensive unified framework to handle a wide array of HEP data processing problems in a simplified and systematic manner and to help to streamline the motley of (sometimes ad hoc) methods currently in use.

## References

[1] Aseev V N et al. 2011 *Phys. Rev. D* **84** 112003 (*Preprint* arXiv:1108.5034).

[2] Tkachov F V 2006 *Transcending The Least Squares* (arXiv:physics/0604127).

[3] Tkachov F V 2013 *Less is more. Why Oberon beats mainstream in complex applications.* Talk

at ACAT 2013, Beijing.

[4] Atwood D and Sony A 1992 *Phys. Rev. D* **45** 2405.

[5] Pretz J 1998 *Reducing the statistical error on the g–2 frequency using a weighting procedure*. Muon g–2 Note No. 326.

[6] Goldblatt R 2006 *Topoi: The Categorial Analysis of Logic*. Dover.

[7] Vilotte J-P 2013 *Data-intensive High Performance Computing and Analysis in Solid Earth Sciences*. Talk at ACAT 2013, Beijing.

[8] Tkachov F V 2002 *Part. Nucl. Lett.* **111** 28-35 (*Preprint* arXiv:physics/0001019).

[9] Bhat P C, Prosper H, Snyder S S 1998 *Int. J. Mod. Phys. A* **13** 5113-5218 (*Preprint* arXiv:hep-ex/9809011).

[10] Tkachov F V, 2002 *Int. J. Mod. Phys. A* **17**, 2783-2884 (arXiv:hep-ph/9901444).

[11] Grigoriev D Yu, Jankowski E, Tkachov F V 2003 *Phys. Rev. Lett.* **91** 061801 (*Preprint* arXiv:hep-ph/0301185).

[12] Tkachov F V 2004 *Nucl. Instrum. Meth. A* **534** 274-278 (*Preprint* arXiv:physics/0401012).