The definition of jets

Fyodor V. Tkachov
Institute for Nuclear Research of Russian Academy of Sciences
Moscow 117312 RUSSIA

A criterion for finding a configuration of jets for a given multiparticle event:
- is derived uniquely from minimum formal input;
- guarantees minimal distortion of physical information in the transition from particles to jets;
- is optimally stable with respect to data errors;
- does not restrict geometric shape of jets;
- allows jets to share particles;
- amounts to finding a matrix via minimization of a simple function;
- does not otherwise restrict algorithmic implementations;
- admits the conventional schemes as simple, perhaps imperfect (sic), heuristics.

Introduction 1

Jet finding algorithms are an important tool in high-energy physics [1], and the problem of quantitative description of the structure of multi-hadron final states remains at the focus of physicists' attention, e.g., in connection with near threshold pair production of particles that decay into jets (such as W^+W^- and $t\bar{t}$; for a review of the issues involved see e.g. [2]). This paper continues the systematic investigation of quantitative description of multijet structure from first principles of physical measurements undertaken in [3]–[5], having in view physical problems of precision measurements type such as studies of the Standard Model rather than phenomenological studies of jet properties. The purpose is to complete the analysis of [4] in regard of jet algorithms and derive a jet finding criterion which is optimal in a well-defined sense, surprisingly simple, and rather rich in implications for the understanding of conventional jet finding algorithms. The focus here on the theory of the criterion, i.e. the derivation and the underlying principles, and how it relates to the conventional algorithms. Software implementation issues are best to be studied in the context of concrete problems.

First in Sec. 2 the criterion is derived in a general form independent of a concrete kinematical situation using only very simple mathematics and developing an idea already touched upon in [4]. In Sec. 3 the remaining ambiguities are fixed for the cases of spherical ($e^+e^- \to$ hadrons) and cylindrical (hadronic collisions) geometries from considerations of momentum conservation and kinematical simplicity, so that there remains no arbitrariness to talk about. Sec. 4 discusses the new criterion with an emphasis on its implications for the understanding of conventional jet finding algorithms.

Notations follow [4] but this paper is self-contained as far as the derivation of the jet definition is concerned. The necessary meta-level propositions of the theory of [4] are summarized in Sec. 2.4 To include a quantity of verbal comments seemed necessary because the theory develops a purely "kinematic" viewpoint which differs from rather widely spread implicit "dynamic" interpretations of jet finding algorithms; because the exposition in ref. [4] may be too technical; and because the result we present here seems to deserve to be well understood in a general perspective.

Jet configuration as approximate description of multiparticle event 2

Abstracting inessential details, a jet finding algorithm transforms a multiparticle event into a configuration of jets. We need to interpret this in a systematic and explicit manner, and to decide what exactly such a transformation is intended to be good for. I consider it important to begin by adopting a minimalistic viewpoint and focus solely on the "calorimetric" information about events. In particular, neither 3- nor 4-momenta for particles are defined until Sec. 3.

Notations. Events and calorimetric detectors 2.1

Let P be the event as seen by an ideal calorimetric detector installation. Then P is a collection of "particles" which will be enumerated using the labels a, b. The a-th particle is represented by its energy E_a (normalized by the total energy of the event: $E_a \leftarrow E_a(\sum_b E_b)^{-1}$) and direction $\hat{\boldsymbol{p}}_a$. Directions can be represented differently (e.g. by φ and θ; by a unit 3-vector; by a point on the unit cylinder; etc.), but all the reasoning until Sec. 3 is independent of the representation. However, the angular distance between two directions, $|\hat{\boldsymbol{p}}-\hat{\boldsymbol{q}}|$, is well-defined irrespective of the chosen representation.

Formally the event **P** is represented as follows:

$$\mathbf{P} = \{E_a, \hat{\boldsymbol{p}}_a\}_a \,. \qquad 2.2$$

The total number of particles in **P** will be denoted as $N(\mathbf{P})$.

An ideal calorimetric detector module is described by its acceptance function f.[i] By definition, for each event **P**, such a module yields the number given by the following expression:

$$\textstyle\sum_a E_a\, f(\hat{\boldsymbol{p}}_a) \equiv \{f, \mathbf{P}\} \,. \qquad 2.3$$

Such expressions will be called <u>*basic shape observables*</u>.

A simple but key proposition is this:

> Knowledge of values of all basic shape observables (expressions 2.3 with all possible f) is equivalent to the knowledge of the event **P**. Therefore, the collection of values of basic shape observables can be regarded as the entire information content of the event **P**.

In point of fact, this proposition directly corresponds to the experimental procedures. Indeed, in real life one deals with a finite collection of calorimetric modules f_n, and with the corresponding finite collection of numbers $\{f_n, \mathbf{P}\}$ for each event **P**. These numbers constitute the experimentally measured approximation to the information content of **P**. If the angular size

[i] f must be, as a minimum, continuous. We will actually impose stronger restrictions which a competent mathematician would easily restore and a physicist won't care about.

of f_n (= the angular region in which $f_n \neq 0$) is sufficiently small for one's purposes, f_n is represented by a direction $\hat{\mathbf{p}}_n$, and we come back to 2.2.

For completeness' sake, note that the values 2.3 can be restored from the collection $\{f_n, \mathbf{P}\}$ bypassing the representation 2.2: it is sufficient to represent an arbitrary f as $f \approx \sum_n c_n f_n$ with some c_n.

By identifying the information content of the event **P** with the collection of expressions 2.3, not only have we not deviated from the experimental reality but we have, in a sense, come closer to it (by allowing finite angular resolutions), at the same time giving it a systematic form convenient for quantitative analysis.

Metaphysics of shape observables 2.4

The meta-level comments below are intended to give a proper perspective to a formally rather simple construction that will follow. Technical details of the construction are independent of these comments.

Each event contains some physical information which is represented by 2.2 or, equivalently, by the collection of shape observables 2.3. This kind of information may be called *raw data*. The ultimate purpose of data processing is to extract *interpreted physical information* such as Lagrangian parameters of the Standard Model.

For instance, to extract the value of the strong coupling $\overline{\alpha}_S$, one introduces observables such as the standard thrust (whatever it is) or the fraction of events deemed to "have 3 jets" (whatever that means). To measure a mass (e.g. of Z) one would invent an observable whose value is a function (rather than one number as with the thrust) with a peak in it whose position is more or less directly determined by the value of the mass.

The bottom line is, to extract interpreted physical information one needs complex observables to be computed from the raw data. All such complex observables which bring out physics and serve to confront models with experiments represent what we call here *observed physical information*. "Interpreted physical information" is model-dependent whereas the "observed physical information" is not; it is, mathematically, simply all possible functions of the raw data (represented either by 2.2 or 2.3) and as such can be examined mathematically.

It turns out that the transition from raw data to observed physical information can be accomplished in ways which are very different algorithmically but are apparently equivalent physically (e.g. the thrust and the 3-jet fraction, which both are theoretically equivalent as instruments for measurement of $\overline{\alpha}_S$). In high-energy physics, the two such concrete classes of observables are:

(a) the observables defined on the basis of intermediate representation of events in terms of jets;

(b) the shape observables computed in a manner similar to the basic shape observables 2.3 (using multiple sums over particles).

Ref. [3] pointed out that the two classes of algorithms differ in their stability to small effects such as data errors, higher-order corrections, algorithmic variations: the shape observables are stable whereas jet algorithms are sometimes not. Ref. [4] systematically developed this line of thought and demonstrated that for practically any observable obtained via jet algorithms (including, say, invariant mass spectra of di-jets, etc.) one can construct a physically equivalent observable bypassing the intermediate representation in terms of jets and using only the so-called C-observables (generalized shape observables) which are obtained from the basic shape observables 2.3 using only "well-behaved" operations (essentially, linear operations and multiplication). So, the two classes of observables are different languages to express the same content. Given that the description in terms of jets involves intrinsic instabilities (which have been known for a long time under the name of "ambiguities of jet definition"), the results of [4] can be interpreted as follows:

The observed physical information is identified with the C-observables which are constructed from the basic shape observables 2.3 using only "regular" operations such as linear operations and multiplication which do not require explicit knowledge of 2.2.

On the other hand, the observables based on jet algorithms express the same physics, may be (much) simpler, but contain enhanced errors due to their instabilities.

Note that although it is impossible to say anything a priori about *interpreted* physical information because this requires a detailed knowledge of theoretical models for the underlying physics, it proved possible to say some important things about *observed* physical information such as the characterization of observables that are optimally stable with respect to data errors via the property of C-continuity as well as point out wide classes of such observables [4].

Also, the theory of [4] refines the general notion of the "observable" for situations encountered in high-energy physics where very different types of data processing algorithms are possible due to the huge complexity of events. In particular, the use of dynamical considerations for construction of observables ought to respect the restrictions imposed on data processing algorithms by the general principles of measurement.[i]

In other words, the principles of measurement restrict the class of one's observables, whereas dynamical considerations are used to select an observable for a particular purpose; and if one has good enough reasons to go beyond the restrictions (e.g. simplicity of calculations), one ought to be aware of the tradeoffs involved (a loss in precision).

Another deep aspect of generalized shape observables [4] is that they naturally fit into the general formalism of quantum field theory because they are, in essence, multiparticle correlators — and the formal apparatus of multiparticle correlators was developed in quantum field theory and statistical mechanics specifically for the purposes of systematic description of systems with an undefined number of particles (as is the case with multiparticle events in high-energy experiments). This point received an ultimate clarification in [5], where C-observables were directly connected to the fundamental energy-momentum tensor. The significance of the latter fact is that it accurately justifies[ii] the empirical rule that the same observable should be used in data processing and in theoretical calculations because the energy-momentum tensor absorbs all the unknown complexity of confinement and hadronization.

---

[i] One such requirement is that the usual hard cuts used for event selection that abound in experimental procedures ought to be regularized [4] — and, on very general grounds, one has to explain not why one regularizes the cuts but why one does not. (Cf. the error analysis in sec. 2.5 of [4]. Recall in this respect that the commonly used statistical methods such as histogramming were originally developed for low precision applications such as demography and agriculture, not high-precision particle physics experimentation; e.g. the procedures for smoothing conventional histograms found in standard numerical packages are not the same as building histograms with regularized bins in the spirit of e.g. sec. 12.9 of [4]; the former entail a loss of information, the latter enhance it by suppressing errors.)

Perhaps it ought to be considered an element of basic culture in data processing that an event should always be accompanied by a weight appropriately modified as the event passes through stages of data processing (instead of the event being simply kept or dropped). Computer memory is cheap enough that extra four bytes per event should not be a burden, and one can always revert to dichotomic weights — but one never quite knows what one looses precision-wise when one sequentially applies a dozen hard cuts to one's events and loosing a few % in precision at each hard cut. Modest as the bang here may be, on a per buck basis it is certainly greater than with hardware upgrades.

[ii] at least in regard of shape observables; concerning jet finding algorithms see Sec. 4.13.

Jets. A qualitative definition 2.5

The basic heuristic of the jet paradigm is to obtain a simplified description of events **P** with $O(100)$ particles in terms of a simpler "configuration of jets" **Q** — which is only a notation yet — which would preserve maximum physical information carried by **P**. Already an explicit formulation of the latter (usually implicit) expectation puts one on the right track leading directly to the definition 2.6.

However, often implicit is also the idea that via a jet algorithm, one restores the pattern of the underlying hard parton event by, in effect, an inversion of hadronization (this idea is particularly prominent in the recombination algorithms). The idea appears to be limited somewhat in view of:

- a non-deterministic nature of hadronization and the intrinsic instability of jet algorithms;
- radiative corrections which blur the correspondence between partons and jets already at the theoretical level;
- a "collective" nature of hadronization (a colored parton becomes a jet of colorless hadrons via interactions with other colored partons).

There is nothing wrong with a dynamical picture as a clue leading to a useful heuristic (see the discussion in Sec. 4.4). However, it seems important that the purely kinematical issues be clarified prior to and irrespective of dynamical considerations. So we return to the kinematical requirement that **Q** should preserve maximum information about **P**.

Given that (i) the entire information content of the event is contained in the observables 2.3; (ii) all meaningful physical information may in principle be extracted from the basic shape observables bypassing intermediate representations in terms of jets; (iii) the observables 2.3 directly incorporate the correct structure of errors in calorimetric measurements, — we demand that any basic shape observable 2.3 should "see", when confronted with the configuration of jets **Q**, roughly the same value as with the original event **P**:

$$\{f,\mathbf{P}\} \approx \{f,\mathbf{Q}\} \quad \text{for any "detector" } f. \tag{2.6}$$

Of course, **Q** must satisfy some requirements, typically have many fewer jets than **P** has particles.

It is hard to believe but just a little mathematical magic is needed to draw forth from 2.6 into broad daylight a quantitative criterion which so remarkably consolidates the subject that no better name for it could be found than *the mother of all jet finding algorithms*. There we go…

Quantitative implications of 2.6 2.7

The definition 2.6 implies that — if we don't want to introduce new hypotheses about how our detectors f see the yet to be defined jets — configurations of jets **Q** must contain, as a minimum, a numerical content similar to that of **P**, meaning that each jet from **Q** is described in the same way as particles of the initial state, i.e. by its energy and direction. One can always extend **Q** (perhaps, having extended **P** first) with additional information and make the jet algorithm compute this additional information (such as jets' 3-momenta; cf. Sec. 3) along with its main job.

To ease interpretation of formulas, jets in **Q** will be labeled by the index j, and the j-th jet's energy and direction will be denoted as $\mathcal{E}_j, \hat{q}_j$. The number of jets in **Q** is $N(\mathbf{Q})$.

The simplest next step is to find jet configurations **Q** which minimize the errors in 2.6, i.e. the quantities

$$\left|\{f,\mathbf{P}\}-\{f,\mathbf{Q}\}\right| = \left|\sum_a E_a f(\hat{\boldsymbol{p}}_a) - \sum_j \mathcal{E}_j f(\hat{\boldsymbol{q}}_j)\right|. \tag{2.8}$$

Our key technical result will be a very simple estimate for 2.8 in which the dependence on f is factorized from the dependence on **P** and **Q**, (see below Eq. 2.24).

Recombination matrix z_{aj} 2.9

To construct a bound for 2.8, one can only compare the values of f at some $\hat{\boldsymbol{p}}_a$ with its values at some $\hat{\boldsymbol{q}}_j$. But which $\hat{\boldsymbol{p}}_a$ to compare with which $\hat{\boldsymbol{q}}_j$ may not be decided from a priori considerations because the dependence on energies plays a central role in the present context and must be incorporated into the analysis from the very beginning.

Introduce the *recombination matrix* z_{aj} (heuristically interpreted as the fraction of a-th particle's energy that goes into the j-th jet; cf. 2.21). Impose the following restrictions on z_{aj}:

$$\bar{z}_a + \sum_j z_{aj} = 1 \quad \text{for any } a, \tag{2.10}$$

$$\bar{z}_a \geq 0, \quad z_{aj} \geq 0 \quad \text{for any } a, j. \tag{2.11}$$

One can see from the derivation that allowing negative values for the matrix elements does not expand the eventual range of options.

- Allowing fractional values for z_{aj} agrees with the physical picture of production of colorless hadrons via interaction of the underlying colored quarks and gluons.
- Allowing non-zero $\bar{z}_a$ corresponds to a finite *energy* resolution and undetected particles (the so-called missing energy).

With the recombination matrix, rewrite 2.8 as follows (the first line is an identical transformation of 2.8, which explains the restriction 2.10):

$$\left|\sum_a \bar{z}_a E_a f(\hat{\boldsymbol{p}}_a) + \sum_a\left(\sum_j z_{aj}\right) E_a f(\hat{\boldsymbol{p}}_a) - \sum_j \mathcal{E}_j f(\hat{\boldsymbol{q}}_j)\right|$$
$$\leq \sum_a \bar{z}_a E_a \left|f(\hat{\boldsymbol{p}}_a)\right| + \sum_j \left|\sum_a z_{aj} E_a f(\hat{\boldsymbol{p}}_a) - \mathcal{E}_j f(\hat{\boldsymbol{q}}_j)\right|. \tag{2.12}$$

One sees why we split particles into fragments corresponding to jets rather than vice versa: we target situations with fewer jets than particles, so it is desirable to arrange cancellations between as many terms as possible in the inner sum, and to minimize the number of positive terms in the outer sum.

Estimating the effects of missing energy 2.13

The first sum on the r.h.s. of 2.12 can be estimated by the following bound:

$$\sum_a \bar{z}_a E_a \left|f(\hat{\boldsymbol{p}}_a)\right| \leq C_{f,1} \sum_a \bar{z}_a E_a. \tag{2.14}$$

The sum on the r.h.s. is a interpreted as the total undetected energy fraction of the event. The constant $C_{f,1}$ is the maximal value of $|f|$ over all directions.

Mathematically also possible are other forms of bounds, but they are not useful here. For instance, a bound in terms of $\max_a(\bar{z}_a E_a)$ cannot be physically useful because it is meaningless to compare energies if two particles' directions are close — one can only add them.

Taylor expansion 2.15

To Taylor-expand $f(\hat{\boldsymbol p}_a)$ near $\hat{\boldsymbol q}_j$ is a little tricky, and we proceed as follows. Consider the plane Π which is normal to the unit 3-vector corresponding to $\hat{\boldsymbol q}_j$. Then map the directions to Π: $\hat{\boldsymbol p} \to \hat{\boldsymbol p}^{\Pi}$, so that the angular distances between directions are preserved near $\hat{\boldsymbol q}_j$:

$$|\hat{\boldsymbol p}^{\Pi} - \hat{\boldsymbol q}_j^{\Pi}| = O(|\hat{\boldsymbol p} - \hat{\boldsymbol q}_j|), \quad \hat{\boldsymbol p} \to \hat{\boldsymbol q}_j, \qquad 2.16$$

where the l.h.s. is a euclidean distance in Π. (An example of such a mapping is given in Sec. 3.2.)

Then $f(\hat{\boldsymbol p})$ becomes a function on Π which we denote as $f_{\Pi}(\hat{\boldsymbol p}^{\Pi}) \equiv f(\hat{\boldsymbol p})$. We will use the Taylor expansion in the form of the following inequality:

$$\left| f_{\Pi}(\hat{\boldsymbol p}_a^{\Pi}) - f_{\Pi}(\hat{\boldsymbol q}_j^{\Pi}) - [\hat{\boldsymbol p}_a^{\Pi} - \hat{\boldsymbol q}_j^{\Pi}] \cdot f'_{\Pi}(\hat{\boldsymbol q}_j^{\Pi}) \right| \le \Delta_{aj} C_{f,3}, \qquad 2.17$$

where $\hat{\boldsymbol p}_a^{\Pi} - \hat{\boldsymbol q}_j^{\Pi}$ is a vector in Π and f'_{Π} is the gradient of f_{Π}. The constant $C_{f,3}$ hides maximal values of some combinations of f and its derivatives through second order. The maximum is taken over all directions $\hat{\boldsymbol q}_j^{\Pi}$ because we will deal with a sum over unspecified $\hat{\boldsymbol q}_j^{\Pi}$.

The only properties which we require the factor Δ_{aj} to have are that it is a monotonic function of the angular distance $|\hat{\boldsymbol p} - \hat{\boldsymbol q}_j|$, and it is such that

$$\Delta_{aj} = O(|\hat{\boldsymbol p} - \hat{\boldsymbol q}_j|^2), \quad \hat{\boldsymbol p} \to \hat{\boldsymbol q}_j. \qquad 2.18$$

It may otherwise be any monotonic function of the angular distance $|\hat{\boldsymbol p} - \hat{\boldsymbol q}_j|$ satisfying 2.18. A modification of Δ_{aj} within these bounds is, roughly speaking, compensated for by an appropriate change of $C_{f,3}$. This observation effectively decouples the form of the r.h.s. of 2.17 from the concrete choice of the mapping Π.

The result 2.17 can be used to estimate the second term on the r.h.s. of 2.12 (add and subtract terms as needed to apply 2.17). Take into account the fact that the values of f at different points are in general independent, so the corresponding expressions have to be bounded independently. Obtain the following upper bound for the second sum on the r.h.s. of 2.12:

$$C_{f,4}\Big(\sum_j \Big|\sum_a z_{aj} E_a - \mathcal{E}_j\Big|\Big) + C_{f,5}\Big(\sum_j \Big|\sum_a z_{aj} E_a [\hat{\boldsymbol p}_a^{\Pi} - \hat{\boldsymbol q}_j^{\Pi}]\Big|\Big) + C_{f,3}\Big(\sum_{j,a} z_{aj} E_a \Delta_{aj}\Big). \qquad 2.19$$

Minimizing 2.19 2.20

The task is to minimize 2.19 using the freedom to choose $\mathcal{E}_j$, $\hat{\boldsymbol q}_j^{\Pi}$, z_{aj}. The arbitrariness associated with Π, and Δ_{aj} will require additional consideration to be eliminated (Sec. 3).

The different f-dependent coefficients are such that they vary independently for arbitrary functions f, so the corresponding expressions have to be treated independently. We are going to suppress the first and second terms.

The first term is suppressed if

$$\mathcal{E}_j = \sum_a z_{aj} E_a \quad \text{for each } j. \qquad 2.21$$

This fixes $\mathcal{E}_j$ in terms of z_{aj}. There are no ambiguities here, and this restriction is naturally interpreted as energy conservation in the formation of jets.

The second term is suppressed if

$$\mathcal{E}_j \hat{\boldsymbol q}_j^{\Pi} = \sum_a z_{aj} E_a \hat{\boldsymbol p}_a^{\Pi} \quad \text{for each } j, \qquad 2.22$$

where we have used 2.21. This determines $\hat{\boldsymbol q}_j$ (via $\hat{\boldsymbol q}_j^{\Pi}$) in terms of z_{aj}, but notice the arbitrariness due to arbitrary Π. Anyhow:

> The two relations 2.21 and 2.22 fix the parameters of jets in terms of the recombination matrix z_{aj} which, therefore, is the fundamental unknown in this scheme.

The final estimate 2.23

Using 2.21 and 2.22 and redefining the f-dependent constants, we find the following simple estimate for 2.19:

$$\left| \{f, \mathbf{P}\} - \{f, \mathbf{Q}\} \right| \le C_{f,1} \sum_{j,a} z_{aj} E_a \Delta_{aj} + C_{f,2} \sum_a \bar{z}_a E_a. \qquad 2.24$$

Lastly, recall that only linear operations and multiplications are needed to construct generalized shape observables (C-observables) of [4] which represent all possible physically meaningful observables. We immediately conclude: **P**

> For any C-observable **H**, the approximation error induced by the transition $\mathbf{P} \to \mathbf{Q}$ with the energies and directions of jets in **Q** determined by 2.21 and 2.22, is bounded by
>
> $$\left| \mathbf{H}(\mathbf{P}) - \mathbf{H}(\mathbf{Q}) \right| \le C_{\mathrm{H},1} \sum_{j,a} z_{aj} E_a \Delta_{aj} + C_{\mathrm{H},2} \sum_a \bar{z}_a E_a. \qquad 2.25$$
>
> where the constants $C_{\mathrm{H},i}$ contain all the dependence on **H**.

The optimal jet definition 2.26

From the estimate 2.25 it follows that the information loss associated with the transition from the event **P** to a configuration of jets **Q** (determined by the recombination matrix z_{aj} via 2.21 and 3.6) are controlled by the following simple functions which depend only on **P** and **Q**:

$$\mathrm{Y}[\mathbf{P},\mathbf{Q}] = \sum_{j,a} z_{aj} E_a \Delta_{aj}, \quad \mathrm{E}[\mathbf{P},\mathbf{Q}] = \sum_a \bar{z}_a E_a. \qquad 2.27$$

The function $\mathrm{Y}[\mathbf{P},\mathbf{Q}]$ is sensitive to presence of clusters of particles in the event **P**, whereas $\mathrm{E}[\mathbf{P},\mathbf{Q}]$ corresponds to undetected particles/soft background.

So, to control the induced errors, one generally needs two positive parameters, which we denote as y_{cut} and $\varepsilon_{\mathrm{miss}}$:

> An optimal configuration of jets **Q** satisfies the following restrictions with a minimal number of jets:
>
> $$\mathrm{Y}[\mathbf{P},\mathbf{Q}] < y_{\mathrm{cut}}, \quad \mathrm{E}[\mathbf{P},\mathbf{Q}] \le \varepsilon_{\mathrm{miss}}. \qquad 2.28$$
>
> Furthermore, the physical meaning of jet finding requires that **Q** should minimize $\mathrm{Y}[\mathbf{P},\mathbf{Q}]$ while saturating $\mathrm{E}[\mathbf{P},\mathbf{Q}]$.

The parameter $\varepsilon_{\mathrm{miss}}$ has a simple and useful meaning of the level of "soft background" above which jets are observed.

Fine-tuning the criterion 3

The jet definition we arrived at in the preceding section contains an arbitrariness which can be fixed by simple additional considerations: momentum conservation and conformance to relativistic kinematics, namely, the hypothesis of effective masslessness of final state hadrons in the high-energy limit. Then with each energy-direction pair $E_a, \hat{\boldsymbol{p}}_a$ one associates a massless Lorentz 4-vector $p_a, p_a^2 = 0$. It will prove convenient to introduce the following 4-vector naturally interpreted as the jet's physical 4-momentum:

$$p_j = \sum_a z_{aj} p_a \,, \qquad p_j^0 \equiv \mathcal{E}_j \,. \qquad 3.1$$

Spherical geometry ($e^+e^- \to$ hadrons) 3.2

Here one emphasizes spherical symmetry. The directions $\hat{\boldsymbol{p}}$ are interpreted as points of the unit sphere, i.e. unit 3-vectors: $\hat{\boldsymbol{p}}^2 = 1$. Then the pair $E_a, \hat{\boldsymbol{p}}_a$ corresponds to the 4-momentum p_a with the energy component $p_a^0 = E_a$ and the 3-momentum component $\boldsymbol{p}_a = E\hat{\boldsymbol{p}}_a$.

3-momentum conservation 3.3

We must choose a mapping of the unit sphere to the plane Π which is normal to $\hat{\boldsymbol{q}}_j$. A simple choice is the stereographic projection from the point $-\hat{\boldsymbol{q}}_j$:

$$\hat{\boldsymbol{p}}_a^{\Pi} = \hat{\boldsymbol{p}}_a + t_{aj}\left(\hat{\boldsymbol{p}}_a + \hat{\boldsymbol{q}}_j\right) = \hat{\boldsymbol{p}}_a + O(\theta_{aj}^2), \qquad 3.4$$

where $t_{aj} = (1-c)/(1+c)$ with $c = \hat{\boldsymbol{p}}_a \cdot \hat{\boldsymbol{q}}_j$. Then Eq. 2.22 is rewritten as

$$\hat{\boldsymbol{q}}_j \mathcal{E}_j = \boldsymbol{p}_j + \sum_a z_{aj} E_a t_{aj}\left(\hat{\boldsymbol{p}}_a + \hat{\boldsymbol{q}}_j\right) = \boldsymbol{p}_j + O(E_a\theta_{aj}^2). \qquad 3.5$$

where $\boldsymbol{p}_j$ is the jet's physical 3-momentum (the space-like component of 3.1).

The arbitrariness in the choice of the mapping Π manifests itself through the terms $O(E_a\theta_{aj}^2)$ in 3.5. A natural choice would be to drop those terms altogether but then one would have to impose a correct normalization on $\hat{\boldsymbol{q}}_j$:

$$\hat{\boldsymbol{q}}_j = \boldsymbol{p}_j / |\boldsymbol{p}_j| \,. \qquad 3.6$$

The direct normalization here cannot take one outside the $O(E_a\theta_{aj}^2)$ arbitrariness in 3.5 (because ensuring a correct normalization of $\hat{\boldsymbol{q}}_j$ was part of the job of the $O(E_a\theta_{aj}^2)$ terms). This can also be verified directly.

Fixing Δ_{aj} 3.7

The only natural choice for Δ_{aj} is this:

$$\Delta_{aj} = 1 - \hat{\boldsymbol{p}}_a \cdot \hat{\boldsymbol{q}}_j \equiv 1 - \cos\theta_{aj} \equiv \tfrac{1}{2}[\hat{\boldsymbol{p}}_a - \hat{\boldsymbol{q}}_j]^2 \equiv E_a^{-1} p_a \tilde{q}_j \,, \qquad 3.8$$

where

$$\tilde{q}_j = (1, \hat{\boldsymbol{q}}_j) \,, \qquad \tilde{q}_j^2 = 0 \,, \qquad 3.9$$

is the light-like Lorentz vector with unit energy associated with the jet's direction $\hat{\boldsymbol{q}}_j$. (It can be called *jet's 4-direction*.) Then

$$\mathrm{Y}[\mathbf{P},\mathbf{Q}] = \sum_j \mathrm{Y}_j[\mathbf{P},\mathbf{Q}] \,, \qquad \mathrm{Y}_j[\mathbf{P},\mathbf{Q}] = p_j \cdot \tilde{q}_j \,, \qquad 3.10$$

where the r.h.s. contains only Lorentz scalar products. Note that in this kinematics $p_j \cdot \tilde{q}_j = \mathcal{E}_j - |\boldsymbol{p}_j|$ but the covariant form 3.10 is more general, as we are going to see shortly.

Cylindrical geometry (hadron-hadron collisions) 3.11

Here one direction (the beam axis) is singled out, and one emphasizes invariance with respect to Lorentz boosts along the beam axis. So one parameterizes directions $\hat{\boldsymbol{p}}$ in terms of a unit vector $\hat{\boldsymbol{p}}^{\perp}, (\hat{\boldsymbol{p}}^{\perp})^2 = 1$ which is orthogonal to the axis (equivalently, by the azimuthal angle $0 \le \varphi < 2\pi$) and the (pseudo-) rapidity $\eta = \ln\cot(\theta/2)$, $-\infty < \eta < +\infty$, which corresponds to a cylindrical geometry. (The polar angle θ is measured from the beam axis.) Then a light-like Lorentz 4-vector p is represented as

$$p = E_{\perp}(\cosh\eta,\ \sinh\eta,\ \hat{\boldsymbol{p}}^{\perp}) \,, \qquad 3.12$$

where $E_{\perp} = E\sin\theta$ is the so-called transverse energy, $E_{\perp}\cosh\theta = p^0$, and $E_{\perp}\sinh\theta$ is the projection of the 3-momentum to the beam axis.

In this case the convention is to replace $E \to E_{\perp} = E\sin\theta$ in all the formulas related to jet definition. Then a reasoning similar to the spherically symmetric case leads one to the following results.

The j-th jet's transverse direction $\hat{\boldsymbol{q}}_j^{\perp}$ is determined similarly to 3.6 from conservation of transverse momentum:

$$\hat{\boldsymbol{q}}_j^{\perp} = \boldsymbol{p}_j^{\perp} / |\boldsymbol{p}_j^{\perp}| , \qquad 3.13$$

with $\boldsymbol{p}_j^{\perp}$ taken from 3.1. For the jet's rapidity one has

$$\eta_j = \sum_a z_{aj} E_a \eta_a \,. \qquad 3.14$$

For Δ_{aj} there is the following simple choice (this structure appeared in [7] in the context of conventional jet algorithms):

$$\Delta_{aj} = \cosh(\eta_a - \eta_j) - \cos(\varphi_a - \varphi_j) \,. \qquad 3.15$$

Then, surprisingly, one recovers 3.10 with

$$\tilde{q}_j = (\cosh\eta_j, \sinh\eta_j, \hat{\boldsymbol{q}}_j^{\perp}) \,, \qquad \tilde{q}_j^2 = 0 \,, \qquad 3.16$$

also the light-like Lorentz vector associated with the jet's direction — specified in this case by the rapidity η_j and the transverse direction $\hat{\boldsymbol{q}}_j^{\perp}$ — and also with unit energy — but in this case it is the transverse energy!

Summary 3.17

Finding the configuration of jets $\mathbf{Q} = \{\mathcal{E}_j, \hat{\boldsymbol{q}}_j\}_j$ for a given event $\mathbf{P} = \{E_a, \hat{\boldsymbol{p}}_a\}_a$ is equivalent to finding the recombination matrix z_{aj} (2.10) which determines jets' parameters via 3.6 (or 3.13 and 3.14) and 3.1 by minimizing $\mathrm{Y}[\mathbf{P},\mathbf{Q}]$ according to 2.28. $\mathrm{Y}[\mathbf{P},\mathbf{Q}]$ is given by 3.10 with the light-like Lorentz vectors $\tilde{q}_j$ (= "jet's 4-direction") given by 3.9 (or 3.16).

Understanding the jet definition 4

□ 4.1

That the criterion 2.28+3.10 is fully constructive is rather wonderful given that it was derived in a straightforward fashion from the seemingly innocent (to the point of appearing meaningless) Eq. 2.6 which, however, only accurately expresses a fundamental idea implicit in the jet paradigm — that the configuration of jets inherits physical information of the corresponding event — an idea so basic that physicists never stopped to ponder it the way it deserves.

The hardest[i] part of all this is, of course, to understand that

- the information content of an event is entirely contained in the collection of basic shape observables 2.3 (see the simple arguments thereafter), and that
- physical information can be extracted from such basic observables via, essentially, only algebraic operations applied to such observables as a whole [4] — which way of extracting physical information is guaranteed not to clash with the specific structure of errors in the data from calorimetric detectors.

Such a description of the event's physical information is in fact *closer to the experimental reality* than the alternative representation in terms of particles', Eq. 2.2, because the latter is based on an interpretation that the event consists of a finite number of point particles.

The latter assumption is actually not necessary for our formalism: just treat the particles' labels a, b as continuous parameters, and replace all sums over a, b by integrals. The recombination matrix z_{aj} then becomes a finite set of functions.

Similarly, the ansatz for the jet configuration $\mathbf{Q}$ that the number of jets should be finite is also not, in general, necessary. This could be an interesting option for a phenomenological description of hadronization but developing the formalism in that direction would take us somewhat too far from the central result of this paper.

Related observables 4.2

Closely related to the criterion is the following monotonically decreasing numeric sequence:

$$\mathbf{J}_m^{\text{opt}}(\mathbf{P}) = \min_{N(\mathbf{Q}')=m} \Upsilon[\mathbf{P},\mathbf{Q}'] \geq 0 . \qquad 4.3$$

It is analogous to the jet-number discriminators $\mathbf{J}_m(\mathbf{P})$ constructed in [3], [4].

Similar observables were earlier introduced on the basis of conventional algorithms [6].

Remarks on implementation 4.4

Computationally, the problem of finding jets in this formulation reduces to finding the recombination matrix z_{aj} which minimizes $\Upsilon[\mathbf{P},\mathbf{Q}]$ within the restrictions 2.10, 2.11. This is rather similar to the classical problem of linear programming for which a vast theory exists (e.g. [8]) where one can borrow ideas for new, efficient, or fancy implementations, which technical subject being well beyond the scope of this Letter, I only make some initial observations.

In general, for events which lit up $O(100)$ detector cells, the huge dimensionality of the minimization problem can make a naïve search of the minima computationally prohibitive, whence the need for good heuristics such as the conventional jet finding algorithms.

A typical implementation of the minimization procedure would contain three elements:

(i) an initial analysis of the configuration (e.g. identification of energetic particles);

(ii) fast but not necessarily always perfect heuristics for finding candidate minima (e.g. recombinations around the energetic particles);

(iii) slower but higher-quality algorithms for fine-tuning the candidate minima and for handling exceptional cases in which the regular heuristics fail.

One liberating consequence is that a concrete implementation of the minimum-finding algorithm is of no consequence whatever (physical or other) provided it yields the optimal jet configuration with required precision. Thus different groups of physicists are free to explore their favorite algorithms — from the simplest gradient method for fastest theoretical computations to cone, recombination, neural, genetic, Danzig's [8], equidomoidal [12], … algorithms — as long as they minimize the same criterion and control approximation errors well in doing so.

□ 4.5

The criterion seems to prefer configurations with z_{aj} equal to exactly 0 or 1 in some simple examples — too simple, however, to allow a generalization. I mention this because even limited information of this kind could be very useful for enhancing efficiency of minimization algorithms (cf. the simplex algorithm [8]). However, with overlapping jets and/or considerable soft background, the option of sharing particles between different jets, which corresponds to fractional values of the recombination matrix z_{aj}, may prove useful.

Cruder alternatives 4.6

Useful for faster search may be cruder versions of the criterion. Two such versions were found in [4] (missing the more fundamental form 2.27 which leads to the remarkable Eq. 3.10) and correspond to replacing $\Upsilon_j[\mathbf{P},\mathbf{Q}]$ by the following expressions which are independent of the jet direction. One is:

$$\overline{\Upsilon}_j[\mathbf{P},\mathbf{Q}] = \max_{a,b}(\hat{p}_a-\hat{p}_b)^2 \times \sum_a z_{aj} E_a \mathcal{E}_j^{-2} \left(\mathcal{E}_j - z_{aj}E_a\right)^2 . \qquad 4.7$$

This is actually more convoluted than Eq. 3.10 but it leads to both Eq. 4.8 and Eq. 4.10 which are useful for comparison with cone and recombination algorithms, respectively.

The other version is

$$\overline{\overline{\Upsilon}}_j[\mathbf{P},\mathbf{Q}] = \max_{a,b}(\hat{p}_a-\hat{p}_b)^2 \times \mathcal{E}_j . \qquad 4.8$$

The two versions are increasingly cruder:

$$\Upsilon_j \leq \overline{\Upsilon}_j \leq \overline{\overline{\Upsilon}}_j . \qquad 4.9$$

For just two whole particles in a jet, Eq. 4.7 becomes

$$\overline{\Upsilon}_j[\mathbf{P},\mathbf{Q}] = \frac{E_a E_b}{\mathcal{E}_j} \times \left[1-\cos\theta_{ab}\right]. \qquad 4.10$$

This is exactly the geometric mean of the JADE [9] and Geneva [10] recombination criteria. (Concerning the connection with the Durham criterion [11] see [4].)

The Geneva and Durham criteria overestimate 4.10 when recombining pairs of soft particles, which is safe. The JADE

[i] for those used to the conventional way of thinking about jet algorithms.

criterion, on the contrary, *under*estimates the errors and is therefore disfavored.

Connection with cone algorithms 4.11

The expression 4.8 is to be compared with the cone algorithms where no energy factors are traditionally used, which indicates their, theoretically speaking, non-optimality e.g. in the cases where jet energies in an event exhibit a significant variation.

On the other hand, the optimal criterion 2.28 agrees with the heuristic sometimes used by experimentalists, which consists in selecting the most energetic particles and defining jets by narrow cones around such particles. From the viewpoint of 2.28 and 3.10, this allows one to directly focus on configurations in which potentially large contributions of energetic particles to $\mathrm{Y}[\mathbf{P},\mathbf{Q}]$ are absent.

Reducing ambiguities in recombination algorithms 4.12

Interestingly, a similar argument allows one to reduce the arbitrariness in the order of recombinations in the conventional iterative $2 \to 1$ recombination algorithms: If the latter are regarded as heuristics for the optimal criterion 2.28 (cf. 4.4), then they should start with the most energetic particles and combine them with their angular neighbors first. This may not help, however, for configurations other than those with a few narrow jets.

A fallacy behind recombination algorithms 4.13

The Conventional Wisdom holds that one should use exactly the same prescriptions in matching experimental and theoretical calculations. This is absolutely correct for all shape observables (as follows from their expressions in terms of the energy-momentum tensor [5]), and hence for our optimal criterion. (No choice is left in regard of the form of $\mathrm{Y}[\mathbf{P},\mathbf{Q}]$ anyway.)

However, CW is only an acceptable default option as long as there are no scientific alternatives. From the viewpoint of the criterion based on a global minimization, the CW recipe amounts to requiring that one should use exactly the same minimization algorithm in the two cases, which — however plausibly it sounds — appears to be missing the point — to the same extent as the events from the sample one deals with are not all narrow-enough jets.

Indeed, the two minimization problems (a few partons in theoretical calculations vs. $O(100)$ hadrons in experimental data processing) have dimensionalities that differ by two orders of magnitude — a **huge** difference for this kind of problems [8]. A competent calculationist's instinct would be to think thrice before relying on a heuristic which worked in one case (e.g. for low dimensional problems) in the other case (a very high dimensionality): a priori one expects two problems so hugely different to require completely different algorithms. Of course, some a priori knowledge of the underlying dynamics helps, but to what extent?

In particular, a conventional algorithm would fail to descend into the minimum correctly for a certain fraction of events. This effect is not dissimilar to spontaneous symmetry breakdown because it occurs even in absence of data errors — it is determined — rather randomly via the order of recombinations — solely by the structure of the algorithm, including the value of y_{cut} (or the jet radius parameter of the cone algorithms to which the arguments here and below are also applicable.)

The meaning of y_{cut} 4.14

In the described scheme, the "jet resolution parameter" y_{cut} actually controls the *errors* induced in the physical information by the approximate description of events in terms of jets. The CW recipe is then equivalent to a hypothesis that the errors thus induced in experimental and in theoretical calculations are correlated well enough to ultimately cancel each other.

But such a hypothesis does not seem to be completely reasonable in view of the spontaneous-symmetry-breaking-like effect of false minima pointed out above. Also take into account that inclusion of higher-order corrections (effectively done to all PT orders in data) multiplies configurations potentially unstable for jet finding/minimum search. Therefore, it appears that the use of the same conventional jet algorithm for data and for QCD calculations introduces poorly controllable $O(y_{\mathrm{cut}})$ distortions into the final results. The new element here compared with the arguments of [3] is that the instability has a character of spontaneous symmetry breaking, its size is determined solely by the algorithm (and is different for different algorithms) but not by data errors, and it does not go away even with ideally precise data and infinite statistics.

It might be possible to study some of these issues with the standard shape observables for $e^+e^- \to$ hadrons.

The role of $\varepsilon_{\mathrm{miss}}$ 4.15

The parameter $\varepsilon_{\mathrm{miss}}$ has a simple meaning as the level of "soft background" above which jets are observed. On the other hand, it corresponds to a common experimental requirement that jets should comprise at least a given energy fraction (like 95%) of the event. It is formally lifted by setting $\varepsilon_{\mathrm{miss}} \geq 1$.

The parameter $\varepsilon_{\mathrm{miss}}$ does not seem to have analogs in the standard schemes for jet-related observables (results are usually presented as functions of y_{cut} only). This may be regarded as their flaw because the two parameters $\varepsilon_{\mathrm{miss}}$ and y_{cut} emerge as a pair very naturally on physical grounds.

Note that the mechanisms manually built into the Geneva and Durham algorithms to prevent wide "lumps" of soft particles from being recognized by the algorithm as jets, play a role not dissimilar to $\varepsilon_{\mathrm{miss}}$. In the latter case, of course, one controls the effect. Also note that with the optimal definition, the soft background is completely excluded from jet formation, whereas Geneva and Durham would have to do something with that background by way of casting it into jets thus perhaps transforming non-perturbative effects into supposedly perturbative jets.

So, including $\varepsilon_{\mathrm{miss}}$ into analyses along with y_{cut} may help:

(i) reduce hadronization effects;

(ii) ease experimentalists' burden in regard of the missing energy by relegating some of the work to theorists;

(iii) reduce ambiguities of conventional algorithms (especially if used together with the prescription to start recombinations with the most energetic particles so that soft particles at wide angular separations from jets are automatic candidates for the soft background).

Conclusions 4.16

Lastly, no jet finding criterion can completely lift the fundamental limitations of the description of data in terms of jets. This also applies to the optimal criterion 2.28+3.10. The optimal criterion at least clarifies the issues involved and offers some new options for improving upon the conventional algorithms.

It seems most attractive for situations with fuzzy or overlapping jets where the conventional algorithms would exhibit greater instability of results. With the optimal criterion, for instance, one need not worry about which jet a pion should be assigned to — its fractions may end up in different jets depending on the values of recombination matrix elements z_{aj}.

The most important practical conclusions from the above discussion appear to be these two:

(a) On quite general grounds, an explicit inclusion of a cut $\varepsilon_{\text{miss}}$ on the soft background along with y_{cut} into the construction of jet-based observables appears to be obligatory. Such a cut has little, if any, direct motivation from the conventional viewpoint but is natural and necessary within the described framework. The complementary parameters $\varepsilon_{\text{miss}}$ and y_{cut} ought to have a comparable status in jet-based analyses.

(b) Concerning the spontaneous-symmetry-breaking-like $O(y_{\text{cut}})$ distortion of final physical results due to non-optimality of the conventional algorithms, it is of course not always possible to say a priori whether the effect is negligible (the case of predominantly needle-like jets) or not.

On the other hand, it is not clear why one would want to forever live with the mess of all the jet finding criteria out there when it can be eliminated in one clean sweep by simply bringing one's jet finding algorithms into a conformance with the optimal definition.

All one seems to need for that to begin with, would be a straightforward piece of code that verifies optimality of and fine-tunes the output of the old algorithm. Since in a majority of events in jet-related experiments the conventional algorithms should work fine (by definition of jet-related experiments), the minimum-search code need not be very efficient (at least at first). Note that such a code is largely independent of the concrete type of problem and of the jet algorithm one deals with: most old-style algorithms should be reasonably good whose output to fine-tune.

Also, the problem in all cases reduces to finding the recombination matrix z_{aj}, and it should not be difficult to design the code so that only a few algorithmic details depend on the particular kinematics of the collision process one deals with. Then one would only have to adapt the interface of such a subroutine to accept one's data format.

ANYHOW:
IRRESPECTIVE OF ANY APPLICATIONS AND CONCRETE CONCLUSIONS,
THE UNCOVERED JET DEFINITION
SHEDS
— AT LONG LAST —
AN UNWAVERING SCIENTIFIC LIGHT
ON THE OBFUSCATED SUBJECT OF JET FINDING,
WHICH IS THE MORE SATISFYING THAT
A CRITERION AS
STUNNINGLY
ELEGANT
AS

$$\boxed{\textstyle\sum_j (p_j \, \tilde{q}_j) < y_{\text{cut}}}$$

DERIVES UNIQUELY
FROM A VERY MODEST FORMAL INPUT.

References

[1] R. Barlow, Rep. Prog. Phys. 36 (1993) 1067.

[2] V.A. Khoze, preprint CERN-TH/98-176 (1998) [hep-ph/9805505].

[3] F.V. Tkachov, Phys. Rev. Lett. 73 (1994) 2405; Erratum, 74 (1995) 2618 [hep-ph/9901332].

[4] F.V. Tkachov, Int. J. Mod. Phys. A12 (1997) 5411 [hep-ph/9601308].

[5] N.A. Sveshnikov and F.V. Tkachov, Phys. Lett. B382 (1996) 403 [hep-ph/9512370].

[6] S. Bethke, in: *QCD–20 Years Later*. World Scientific, 1993.

[7] S. Catani et al., Nucl. Phys. B406 (1993) 187.

[8] C.H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization Algorithms and Complexity*. Prentice-Hall, 1982.

[9] JADE collaboration (W. Bartel et al.), Z. Phys. C33 (1986) 23.

[10] S. Bethke et al., Nucl. Phys. B370 (1992) 310.

[11] S. Catani et al., Phys. Lett. 269B (1991) 432.

[12] R. Queneau, *Bords,* Hermann, 1963.