INSTITUTE OF PHYSICS PUBLISHING REPORTS ON PROGRESS IN PHYSICS

Rep. Prog. Phys. **67** (2004) 553–622 PII: S0034-4885(04)04034-5

Pattern recognition and event reconstruction in particle physics experiments

R Mankel

DESY, Notkestraße 85, D-22603 Hamburg, Germany

E-mail: Rainer.Mankel@desy.de

Received 28 January 2004
Published 25 March 2004
Online at stacks.iop.org/RoPP/67/553
DOI: 10.1088/0034-4885/67/4/R03

Abstract

This report reviews methods of pattern recognition and event reconstruction used in modern high energy physics experiments. After a brief introduction to general concepts about particle detectors and statistical evaluation, different approaches in global and local methods of track pattern recognition are reviewed with their typical strengths and shortcomings. The emphasis is then shifted to methods which estimate the particle properties from those signals which pattern recognition has associated. Finally, the global reconstruction of the event is briefly addressed.

(Some figures in this article are in colour only in the electronic version)

Contents

	Page
1. Introduction	556
2. Basics	556
2.1. Detector layouts	556
2.1.1. Forward or fixed target geometry	558
2.1.2. Collider detector geometry	559
2.2. Typical tracking devices	560
2.2.1. Linear single-coordinate measurements	560
2.2.2. Radial single-coordinate measurements	560
2.2.3. Stereo angles	562
2.2.4. 3D measurements	563
2.3. Track models and parameter representations	564
2.3.1. Forward geometry	564
2.3.2. Cylindrical geometry	564
2.4. Parameter estimation	565
2.4.1. Least squares estimation	565
2.4.2. The Kalman filter technique	566
2.5. Evaluation of performance	568
2.5.1. The reference set	568
2.5.2. Track finding efficiency	568
2.5.3. Ghosts	569
2.5.4. Clones	569
2.5.5. Parameter resolution	569
2.5.6. Interplay	570
3. Global methods of pattern recognition	570
3.1. Template matching	571
3.2. The fuzzy radon transform	572
3.3. Histogramming	574
3.4. Neural network techniques	578
3.4.1. The Hopfield neuron	578
3.4.2. The Denby–Peterson method	579
3.4.3. Elastic arms and deformable templates	582
4. Local methods of pattern recognition	589
4.1. Seeds	589
4.2. 2D versus 3D propagation	590
4.3. Naïve track following	592
4.4. Combinatorial track following	593
4.5. Use of the Kalman filter	593
4.6. Arbitration	593
4.7. An example for arbitrated track following	594
4.7.1. Algorithm	594
4.7.2. Parameters	595

4.8. Track following and impact of detector design parameters 596
4.8.1. Influence of detector efficiency 596
4.8.2. Effect of detector resolution 596
4.8.3. Influence of double track separation 596
4.8.4. Execution speed 596
4.9. Track propagation in a magnetic field 599
5. Fitting of particle trajectories 599
5.1. Random perturbations 599
5.2. Treatment of multiple scattering 600
5.2.1. Impact parameter and angular resolutions 602
5.2.2. Momentum resolution 604
5.2.3. Effects of fit non-linearity 604
5.2.4. Contributions of different parts of the spectrometer 605
5.2.5. Parameter covariance matrix estimation 605
5.2.6. Goodness of fit 606
5.3. Treatment of ionization energy loss and radiation 607
5.3.1. Ionization energy loss 607
5.3.2. Radiative energy loss 608
5.3.3. Radiation energy loss correction within the magnetic field 611
5.4. Robust estimation 613
6. Event reconstruction 616
6.1. Vertex pattern recognition 616
6.2. Vertex fitting 617
6.3. Kinematical constraints 618
7. Concluding remarks 619
Acknowledgment 619
References 619

1. Introduction

Scientific discovery in elementary particle physics is largely driven by the quest for higher and higher energies, which allow us to delve ever more deeply into the fine structure of the microscopic universe. Higher energies lead, in general, to an increased multiplicity of particles. Since the acceleration of electrons is limited either by synchrotron radiation in the case of storage rings, or by field gradients in the case of linear colliders, multi-TeV energies are in the near future only accessible by accelerating hadrons, the collision of which generates even more particles.

Reconstruction of charged particles from signals of tracking detectors in spectrometers has always shown aspects of a discipline of art, since the variety of experimental setups lead to the development of very diverse pattern recognition methods, which cannot easily be ranked against each other. A general overview has been given in an earlier review [1]. It is remarkable that even today, no generally accepted standard software package exists which performs track finding in a variety of detector setups, a situation which is in marked contrast, e.g., to detector simulation. A new generation of experiments is now emerging in which the track density is so high that success will crucially depend on the power of the reconstruction methods. One example for the development in tracking demands over 15 years is illustrated in figure 1, which shows in direct comparison an event from the experiment ARGUS [2], which took data of e^+e^- collisions in the Υ range in the period 1982–1992, and the ATLAS experiment [3] that is currently under construction and will operate from 2007 onwards with proton collisions at the LHC. The new experiments also require huge computing resources for reconstruction of their data. Since track finding is usually the most time consuming part in reconstruction, the sophistication and economy of pattern recognition methods has considerable impact on the computing effort.

Pattern recognition plays an important rôle in other detector components also, for example, cluster reconstruction in calorimeters, or ring finding in ring imaging Čerenkov (RICH) detectors. It is, however, in track reconstruction where the new generations of experiments pose the most crucial challenges. This paper will, therefore, focus on track reconstruction as well as on related aspects of event reconstruction.

The first of the following sections will provide an introduction to basic detector concepts and tracking devices and summarize mathematical tools for estimating parameters and performance, which will be used later on. The two following sections focus on track pattern recognition using various methods, including applications in several experiments. The next section then concentrates on parameter estimation from particle trajectories, which is—in contrast to track finding—in principle a straightforward mathematical problem, but contains several detailed issues worth mentioning. The last section briefly discusses some track-related aspects of event reconstruction.

2. Basics

This section provides a brief introduction to the basic elements influencing event reconstruction. It is not intended to cover the subject of particle detectors in full detail; instead references to the relevant detector literature (see, e.g., [4–6]) are given.

2.1. Detector layouts

Modern detectors in high energy physics are usually sampling detectors. The detector volume is filled with devices which the particles traverse and in which they leave elementary pieces of

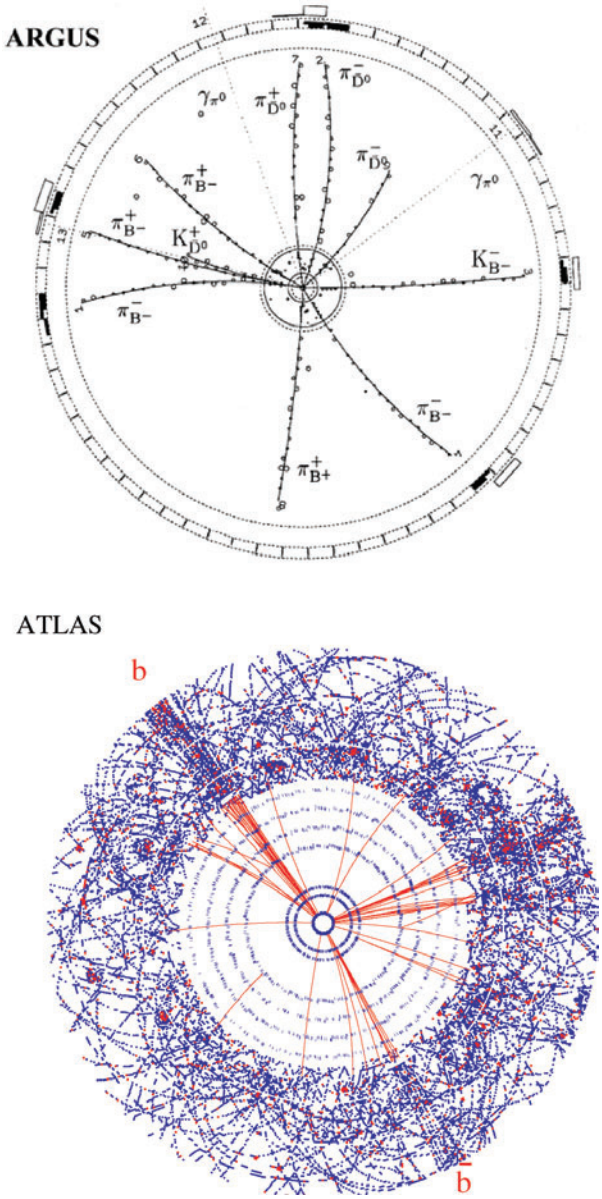

Figure 1. Comparison of event complexity in the experiments ARGUS and ATLAS. The ARGUS event (top) consists of two reconstructed B mesons, one of them being a candidate for the charmless decay $B^- \to K^- 4\pi^\pm$ (from [2]). The ATLAS display (bottom) shows a simulation of an event in the inner detector with a Higgs boson in the decay mode $H^0 \to b\bar{b}$, including the pileup at full LHC luminosity (from [3]).

information, as, e.g., an excitation in a solid-state detector, a primary ionization in a gaseous chamber or an energy deposition in the sensitive volume of a calorimeter. The event record of an experiment consists of the amassed volume of the signals from all particles of an interaction—or possibly even several interactions—which are collated. After sorting out which bits of information are related to the same particle—this process is called pattern recognition—the

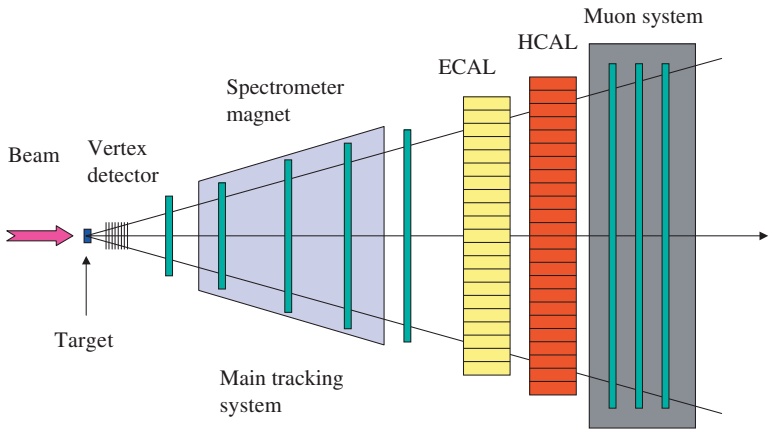

Figure 2. The typical geometry of a forward spectrometer, as used, e.g., in fixed-target setups.

kinematical properties of each particle have to be reconstructed to reveal the physical nature of the whole event.

In general, experiments nowadays strive to record the interaction as a whole, with all (significant) particles produced in the process. This has led to the development of 4π detectors, where almost the whole solid angle region, as seen from the interaction, is covered.

In general, two main concepts have to be distinguished, which will be discussed in the following.

2.1.1. Forward or fixed target geometry. When the interaction is generated by a beam incident on a fixed target, the centre-of-mass system of the participating particles is seen under a strong Lorentz boost, and the emerging particles move within a cone in the forward direction. In this case, the detector set-up must cover this forward cone with the instrumentation, while the part of the solid angle further behind is generally neglected. This scenario is called a forward detector geometry. Similar situations exist where the dynamics of the interaction results in all relevant particles being produced under a huge Lorentz boost, like heavy flavour production at large hadron colliders.

Figure 2 shows schematically a forward detector geometry as is used in fixed target experiments. The event is generated through collision of a beam particle with a nucleus in the target. Because of the momentum of the incident beam particles, the whole event is seen under a Lorentz boost in the beam direction, so that the emerging particles are confined to a cone whose opening angle depends on the typical transverse momenta generated in the interaction, and the size of the Lorentz boost.

The main components of a typical forward spectrometer are described below.

- The vertex detector, which is a precision tracking system very close to the interaction point. Its main purpose is the improvement of track resolution near the interaction point which allows reconstruction of secondary vertices or distinguishing of detached tracks that are used, e.g., for the tagging of heavy flavour decays.
- The spectrometer magnet with the main tracking system, which measures the trajectories of charged particles and determines their momentum and charge sign from the curvature.
- The calorimeter system, which is often split into an electromagnetic and a hadronic part. The calorimeter allows identification of electrons and hadrons by the shower

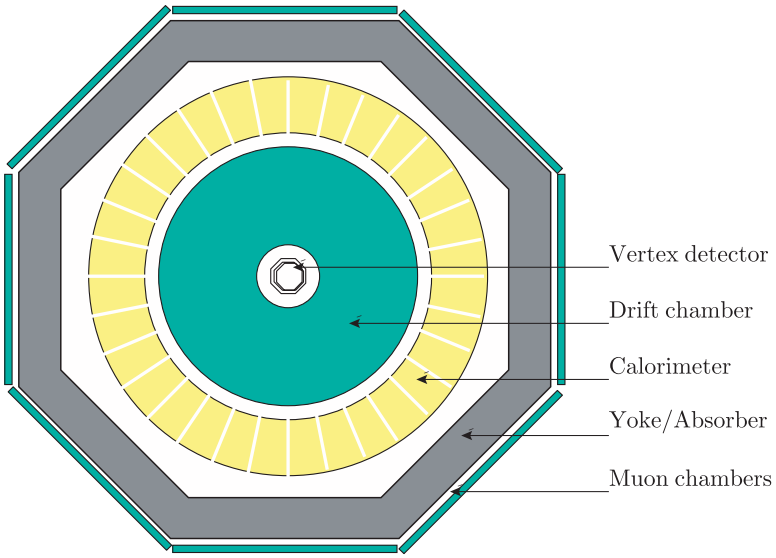

Figure 3. Typical set-up of a collider detector.

energy deposited, and very often provides essential signals for the trigger system. The calorimeter can also measure energies of individual neutral particles, in particular photons, though its actual capability for this task depends strongly on the particle density in the event.

- The muon detector, which consists of tracking devices in combination with absorbers. Only muons are able to traverse the intermediate material, and are then measured in the dedicated tracking layers.

The design of a forward spectrometer is influenced by several factors. The sheer size of the tracking volume depends on the leverage required for the momentum resolution, since at sufficiently high momentum the resolution is inversely proportional to the integral of the magnetic field along the trajectory [7], as will be discussed in more detail in section 5. Depending on the scope of the experiment, further detector components may be introduced to provide particle identification, for example, RICH counters or transition radiation detectors (TRD).

2.1.2. Collider detector geometry. When two beams collide head-on, the centre-of-mass system of the interactions is either at rest or moving moderately. In this case, the detector should try to cover the full solid angle. This beam setup usually leads to cylindrical detector layouts with a solenoid field parallel to the beam axis (figure 3). The cylindrical geometry detector differs in several details from the forward geometry detector:

- The vertex detector requires modules parallel to the beam, at least in the central part of the angular acceptance, often referred to as the barrel part.
- The main tracking system is generally contained in the magnetic field. The coil and yoke of the magnet usually have to be within the detector volume, where the general choice is to have the coil between the drift chamber and the calorimeter, where particles traverse it before their energy is measured in the calorimeter, or to make it large enough to enclose the calorimeter. This may be more costly to build and operate and the field may have adverse effects on the calorimeter itself.

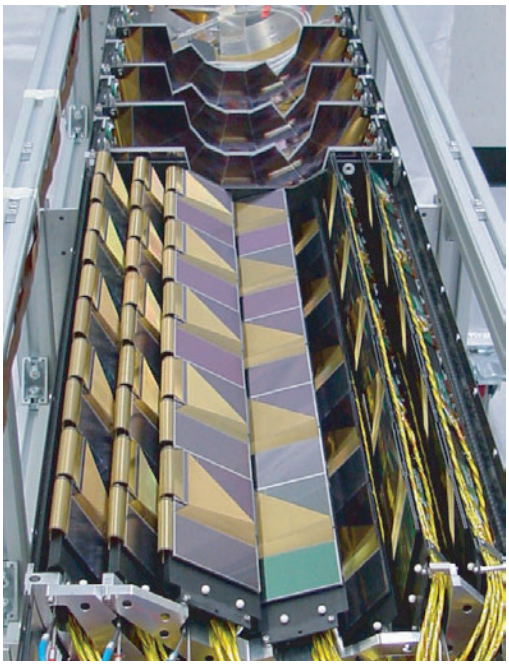

Figure 4. Lower half barrel part of the Zeus micro-vertex detector.

- The calorimeter system now requires barrel and end cap parts to cover the solid angle. One main function of the calorimeter at high energy colliders is the measurement of jets.
- For the muon detector, the yoke of the solenoid lends itself readily as an absorber.

2.2. Typical tracking devices

2.2.1. Linear single-coordinate measurements. One type of tracking device, which is in widespread use measures one coordinate of the particle whose trajectory intersects the device. A good example for this type is the silicon strip detector, which is a semiconductor-based device made of strips typically down to widths of 25 μm. Each strip works like a small diode, with a voltage applied such that the border area is depleted and the resistance is high. A traversing charged particle will then create pairs of electrons and corresponding holes which drift apart under the voltage and can be registered as a pulse. In general, several strips will register a signal under traversal of a particle, and the pulse heights of the participating channels can be evaluated with suitable clustering algorithms, for example, centre-of-gravity based, and determine the location at which the particle has passed. Solid-state detectors are presently the tracking devices with the highest spatial resolution, and they are often installed very close to the interaction region as vertex detectors where they allow or improve the reconstruction of primary and secondary vertices. Another favourable property of solid-state detectors is their resilience against radiation damage. The current limitation is in the size of individual detector modules, which makes them expensive for covering large volumes. Figure 4 shows the micro-vertex detector of the ZEUS experiment [8], prior to its installation in 2001.

2.2.2. Radial single-coordinate measurements. The size of the tracking volumes is important, since momentum measurement requires the particle to traverse a magnetic field, where the

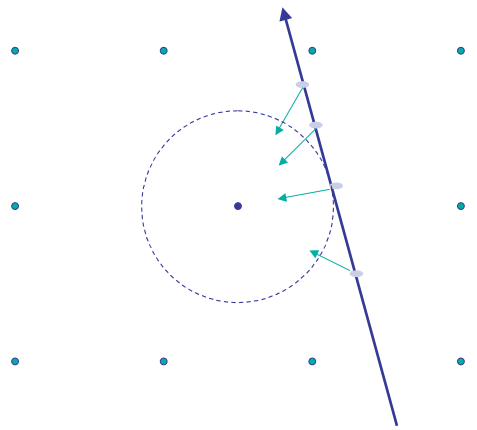

Figure 5. Schematic view of a drift chamber cell. The filled circles indicate wires, with the sense wire in the middle of the cell and the field wires on the outside. The black arrow shows the trajectory of a particle, the grey arrows denote primary ionization charges drifting towards the sense wire.

length of the path provides the leverage that determines the precision of the momentum reconstruction. This is one of the reasons why gaseous chambers, in particular drift chambers, are very commonly employed when large areas have to be covered.

The basic principle of the drift chamber is displayed in figure 5. A drift cell consists of an anode wire in the centre and an arrangement of field wires. The geometry shown is very similar to that in the ARGUS drift chamber [9] (see also figure 34). The drift cell need not be rectangular in shape: in the drift chamber of the BaBar experiment, for example, it is hexagonal [10]. Primary ionization occurs along the path of the particle. The charges drift to the anode wire, where they create a locally confined avalanche of particles within the large electrical field close to the wire. This effect results in a multiplication of the ionization which is called gas amplification. The rising edge of the signal picked up by the anode wire triggers a time-to-digital converter (TDC) which then measures the time until a common stop signal. This allows the measurement of the drift time for those charges that are the first to arrive. In the simplest case, the drift field will be shaped such that the drift velocity is uniform, and the time resolution can be directly transformed into a uniform resolution of the drift distance. In practice, numerous effects can lead to a non-linear drift-time/space relation, and the spatial resolution will depend on the precise location of the traversal of the particle.

Since the time measured by the TDC corresponds to the arrival of the first charges, usually those with the smallest distance to the wire, the drift chamber measures the distance of closest approach of the particle to the wire. In those cases where more than one particle traverses the same drift cell within the same interaction window, in general only the particle closest to the wire is registered. This effect may cause complications for pattern recognition which depend on the degree of occupancy. Another typical property of drift chambers is that the single measurement cannot distinguish on which side of the wire the particle has traversed the cell; this uncertainty is called left–right ambiguity. In the worst case, left–right ambiguity may lead to a mirror track that cannot be distinguished from the real one. Theories have, therefore, been developed on how to design drift chambers such that left–right ambiguity can be resolved in all cases, e.g., the butterfly geometry [11].

Drift in gases is also influenced by magnetic fields. The deviation of the gas drift direction from the vector of the electric field is described by the Lorentz angle. Figure 6 shows an event

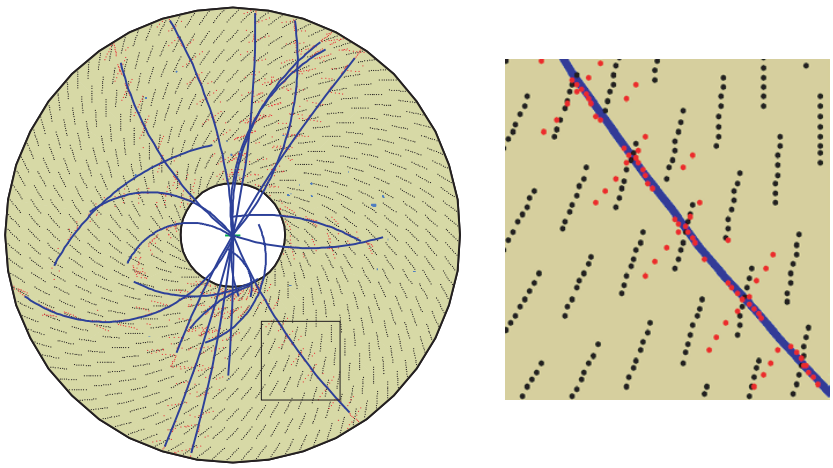

Figure 6. Left: event display from the ZEUS CTD, showing sense wires and reconstructed tracks. Right: closeup around the track in the lower left area. The black dots represent the sense wires, the grey dots indicate the drift distance end points on both sides of the corresponding wire.

display of the central tracking detector (CTD) of the ZEUS experiment, in the view along the beam axis, which has been created using the tool described in [12]. The Lorentz angle in this case is 45°, and it is reflected in the design of the cell structure.

2.2.3. Stereo angles. Devices measuring single coordinates do not provide three-dimensional[1] points on a trajectory, but measure only in a projected space. While such devices can be very economical in the sense that a relatively small number of channels is needed to cover a region at good resolution, 3D information can only be obtained by combining several projections, usually called stereo views. While two views are in principle sufficient to reconstruct the spatial information, the presence of more than one track, in general, leads to ambiguities regarding the assignment of projected information. This is illustrated in figure 7, where two particles are measured in two strip detector views of 0° (x) and -45° (u). Ambiguity in the assignment of the measured hits in the x and u views to each other leads to the reconstruction of two ghost points. This illustrates that, in general, at least three views are necessary to avoid this kind of ambiguities. On the other hand, in special cases of limited track density, the use of only two views may be justified, since in this case the majority of ghosts may be discarded for geometrical reasons. This can already be guessed from figure 7; since the true tracks are well separated, the uppermost ghost combination is already just outside the chamber acceptance of the u view. Such concepts are called all-stereo designs.

An example of a spectrometer that combines several types of single-coordinate measurements is the HERA-B detector [13–15], which is shown in figure 8. The vertex detector (labelled SI) consists of eight superlayers of silicon strip detectors with four different stereo angles. The design of the main tracker is structured into three areas: within the magnet (MC), between the magnet and the RICH (PC) and between the RICH and the calorimeter (TC). It contains 13 superlayers of honeycomb drift chamber modules for the outer area and ten superlayers of micro-strip gaseous chambers (MSGC) for the region close to the beam[2].

[1] The shorthand notations two- (2D) and three-dimensional (3D) will frequently be used in the following.
[2] The layout of tracking stations was modified later with the shift in emphasis away from B physics.

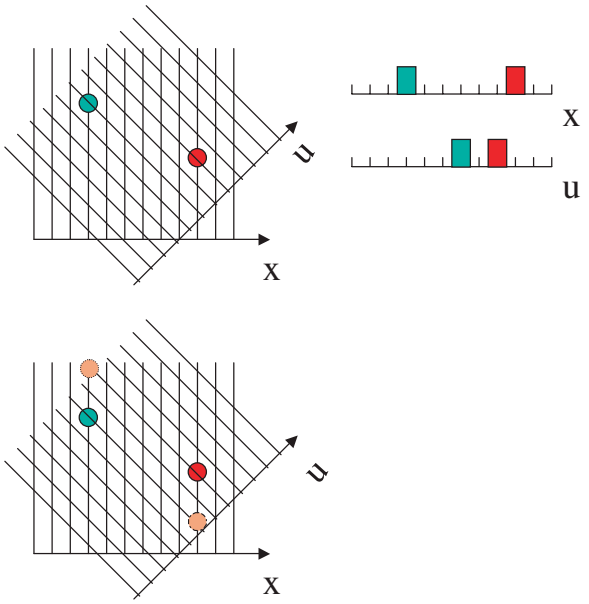

Figure 7. Hit ambiguities with two stereo views.

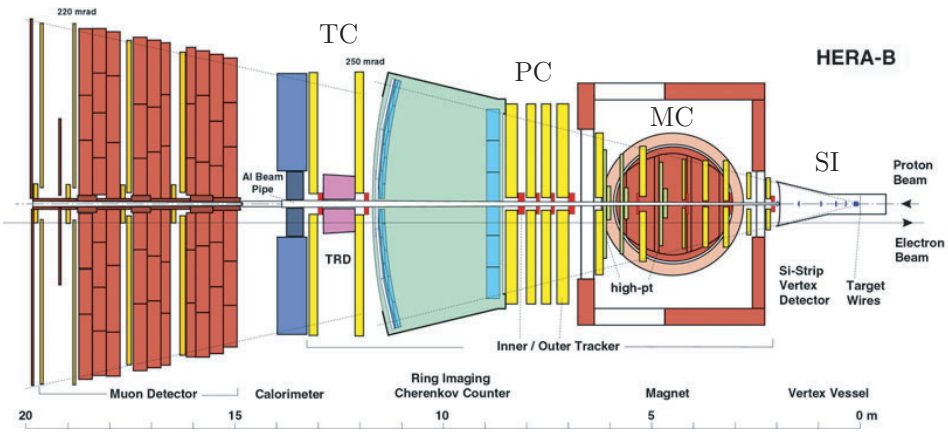

Figure 8. Layout of the HERA-B spectrometer. The labels TC, PC, MC and SI indicate groups of tracking stations that comprise the vertex and main tracking system.

2.2.4. 3D measurements. In general, pattern recognition will benefit considerably if the tracking device itself is able to measure 3D space points. A modern example is solid-state pixel detectors, as for example, the CCD-based vertex detector of the SLD experiment [17], where the pixels have a size of $20 \times 20\,\mu\text{m}^2$. A gaseous detector capable of covering large tracking volumes with 3D measurement is the time projection chamber (TPC). Figure 9 shows the TPC of the STAR experiment [16]. The gas volume itself is free of wires; instead, an axial electric field, produced with the help of a membrane electrode in the middle plane, lets the primary charges drift to the anodes at the end caps, where they are registered, for example, with multi-wire proportional chambers with pad readout. While this provides a direct measurement of the x and y coordinates, the z coordinate is inferred from the time measurement. The

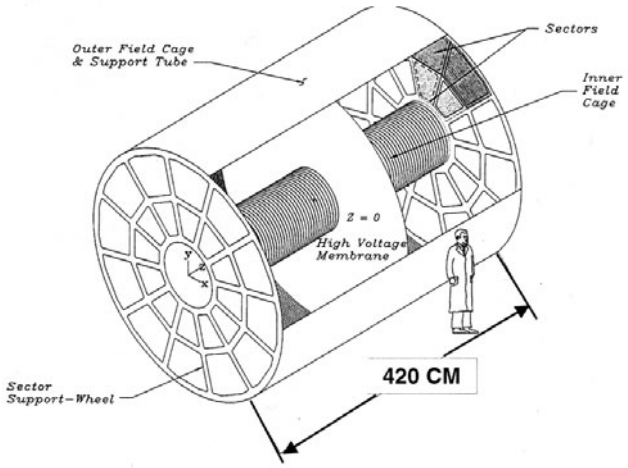

Figure 9. TPC of the STAR experiment (from [16]).

magnetic field is also axial, and plays an important rôle in limiting diffusion effects during the drift.

2.3. Track models and parameter representations

2.3.1. Forward geometry. In the forward geometry, the interaction region lies very often in an area without magnetic field, since the spectrometer magnet is located further downstream. The natural choice of parameters, assuming that the z coordinate points down the spectrometer axis and x and y are the transverse coordinates, is then

x_0, the x coordinate at the reference z_0;
y_0, the y coordinate at the reference z_0;
$t_x = \tan\theta_x$, the track slope in the xz plane;
$t_y = \tan\theta_y$, the track slope in the yz plane;
Q/p, the inverse particle momentum, which is positive or negative depending on the charge;

where z_0 denotes the location of a suitable reference plane transverse to the beam, for example, at the position of the target, or at the nominal interaction point. The slope parameters allow for a convenient transformation of the parameters to a different reference z value, as is needed during vertex reconstruction. In cases of a very homogeneous magnetic field, it may be advantageous to substitute the parameter Q/p by $Q/p_\perp$, where $p_\perp$ is the momentum in the plane transverse to the magnetic field, or by $\kappa = Q/R$, the signed inverse radius of curvature.

2.3.2. Cylindrical geometry. In collider detectors with cylindrical geometry, the magnetic field normally encompasses the whole tracking volume, including the interaction region where the particles are produced. In a homogeneous solenoid field, the particle trajectory will be a helix curling around an axis parallel to the magnetic field. Assuming the z coordinate is oriented along the detector axis, and the radius is given by $r = \sqrt{x^2 + y^2}$, typical track parameters given at a reference value $r = r_0$ may be

ϕ_0, the azimuthal angle where the trajectory intersects the reference radius;
z_0, the z value where the trajectory intersects the reference radius;

ψ_0, the phase angle of the helix at the reference radius intersection, which corresponds to the angle of the tangent at this point;
Q/R, the signed inverse curvature radius of the helix;
$\tan\lambda$, where $\lambda = \arctan p_z/p_\perp$ is the dip angle of the helix.

2.4. Parameter estimation

The estimation of the kinematical parameters of a particle, such as position (or impact parameter), direction of flight and momentum at its point of origin, from spatial measurements along its trajectory is generally referred to as track fitting. We assume at this point that the measurements related to a particle have been correctly identified in the pattern recognition step (which will be discussed in more detail in sections 3 and 4). A very general approach to parameter estimation is the maximum likelihood method, which will not be discussed here in detail; instead, we refer to the textbook literature [18–22]. The maximum likelihood method can take very general distributions of the observed variables into account, for example, exponential distributions as they may occur when decay lengths are measured. However, its application in multi-parameter problems can be very complex, in particular the error analysis. In cases where the distribution of the random variables is Gaussian, at least approximately, the least squares method is generally successful. Since many observables in track reconstruction do at least approximately follow a Gaussian distribution, we will focus on this method in the following.

2.4.1. Least squares estimation. If the trajectory of a particle can be described by a closed expression $f_{\vec{\lambda}}(\ell)$, where $\vec{\lambda}$ stands for the set of parameters, ℓ is the flight path and f the coordinate which could be measured, a set of measurements $\{m_i\}$ with errors $\{\sigma_i\}$ will provide an estimate of the parameters according to the least squares principle

$$X^2 = \sum \frac{(m_i - f_{\vec{\lambda}}(\ell_i))^2}{\sigma_i^2} \stackrel{!}{=} \min. \tag{1}$$

One can easily convince oneself that in the case of normally distributed measurements m_i, the above impression is proportional to the negative logarithm of the corresponding likelihood function, which shows directly the equivalence of the least squares principle and the maximum likelihood principle for this case.

Symbolizing the derivative matrix[3] of f with respect to the parameters as $\boldsymbol{F}$ and the (diagonal) error matrix of the measurements as $\boldsymbol{V} = \mathrm{diag}\{\sigma_i^2\}$, the expression to be minimized is

$$(\vec{m} - F\vec{\lambda})^T V^{-1} (\vec{m} - F\vec{\lambda}) \tag{2}$$

and requiring the derivative to vanish at the minimum leads to the matrix equation

$$F^T V^{-1} \vec{f} = F^T V^{-1} \vec{m}. \tag{3}$$

In the case of a linear problem, $\vec{f} = F\vec{\lambda}$, the above condition can be directly inverted

$$\vec{\lambda} = (F^T V^{-1} F)^{-1} F^T V^{-1} \vec{m} \tag{4}$$

and the estimated parameters are a linear function of the measurements. The matrix $(F^T V^{-1} F)^{-1}$ that needs to be inverted is of the form $N_\lambda \times N_\lambda$ (where N_λ is the number

[3] We denote the derivative matrix as $\partial f/\partial\lambda$, where $(\partial f/\partial\lambda)_{ij} = \partial f_{\vec{\lambda}}(\ell_i)/\partial\lambda_j$.

of parameters describing the particle), and this process is computationally inexpensive. Also, the covariance matrix of the parameter estimate can be directly determined as

$$\mathrm{cov}(\vec{\lambda}) = C_\lambda = (F^T V^{-1} F)^{-1}. \tag{5}$$

The popularity of the least squares method can be attributed to its optimality properties in the linear case:

- the estimate is unbiased, i.e. the expectation value of the estimate is the true value;
- the estimate is efficient, which means that, of all the unbiased estimates which are linear functions of the observables, the least squares estimate has the smallest variance. This is called the 'Gauss–Markov–Theorem'.

Though these properties are strictly guaranteed only for the linear case, they are still retained in most cases where the function $f_{\tilde{\lambda}}$ can be locally approximated by a linear expansion.

The expression X^2 in equation (1) will follow a χ^2 distribution if the function f_λ is (sufficiently) linear and if the measurements m_i follow a normal distribution. This property can be used for statistical tests. In particular, the second condition should always be kept in mind, as its relevance will become apparent later.

2.4.2. The Kalman filter technique. The least squares parameter estimation as described in the previous section requires the global availability of all measurements at fitting time. There are cases when this requirement is not convenient, for example, in real-time tracking of objects, or in pattern recognition schemes which are based on track following, where it is not clear *a priori* if the hit combination under consideration does really belong to an actual track.

The Kalman filter technique was developed to determine the trajectory of the state vector of a dynamical system from a set of measurements taken at different times [23]. In contrast to a global fit, the Kalman filter proceeds progressively from one measurement to the next, improving the knowledge about the trajectory with each new measurement. Tracking of a ballistic object on a radar screen may serve as a technical example. With a traditional global fit, this would require a time consuming complete refit of the trajectory with each added measurement.

Several properties make the Kalman filter technique an ideal instrument for track (and vertex) reconstruction [24–26]. The prediction step, in which an estimate is made for the next measurement from the current knowledge of the state vector, is very useful to discard noise signals and hits from other tracks from the fit. The filter step which updates the state vector does not require inversion of a matrix with the dimensions of the state vector, as in a global fit, but only with the dimensions of the measurement, leading to a very fast algorithm. Finally, the problem of random perturbations on the trajectory, as multiple scattering or energy loss, can be accounted for in a very efficient way. In its final result, the Kalman filter process is equivalent to a least squares fit.

In this paper, the implementation and nomenclature from [25, 27] is used, and these documents are referred to for a more detailed explanation of the Kalman filter method. In this notation, the system state vector at the time k, i.e. after inclusion of k measurements, is denoted by $\tilde{x}_k$ and its covariance matrix by C_k. In our case, $\tilde{x}_k$ contains the parameters of the fitted track, given at the position of the kth hit. The matrix F_k describes the propagation of the track parameters from the $(k-1)$th to the kth hit[4]. For example, in a planar geometry with

[4] We assume at this stage a linear system, so that F_k and H_k are matrices in the proper sense. For treatment of the non-linear case see later.

one-dimensional (1D) measurements and straight-line tracks, the propagation takes the form

$$\begin{pmatrix} x \\ t_x \end{pmatrix}_k = \begin{pmatrix} 1 & z_k - z_{k-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ t_x \end{pmatrix}_{k-1} \tag{6}$$

where a subset of the track parametrization in section 2.3.1 has been used. The coordinate measured by the kth hit is denoted by m_k. In general, m_k is a vector with the dimension of that specific measurement. For tracking devices measuring only one coordinate, m_k is an ordinary number. The measurement error is described by the covariance matrix V_k. The relation between the track parameters $\tilde{x}_k$ and the predicted measurement is described by the projection matrix H_k. In the example in section 2.2.3, the measured coordinate in the stereo view u is

$$H \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\alpha_{\mathrm{st}} & -\sin\alpha_{\mathrm{st}} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{7}$$

with α_{st} as the stereo angle (-45° in the example).

In each filter step, the state vector and its covariance matrix are propagated to the location or time of the next measurement with the prediction equations:

$$\tilde{x}_k^{k-1} = F_k \tilde{x}_{k-1} \qquad C_k^{k-1} = F_k C_{k-1} F_k^T + Q_k \tag{8}$$

and the estimated residual becomes

$$r_k^{k-1} = m_k - H_k \tilde{x}_k^{k-1} \qquad R_k^{k-1} = V_k + H_k C_k^{k-1} H_k^T . \tag{9}$$

Here, Q_k denotes the additional error introduced by process noise, i.e. random perturbations of the particle trajectory, for example, multiple scattering. We will see later (section 5.2) how this treatment works in detail. The updating of the system state vector with the kth measurement is performed with the filter equations:

$$\begin{aligned} K_k &= C_k^{k-1} H^T (V_k + H_k C_k^{k-1} H_k^T)^{-1} \\ \tilde{x}_k &= \tilde{x}_k^{k-1} + K_k (m_k - H_k \tilde{x}_k^{k-1}) \\ C_k &= (1 - K_k H_k) C_k^{k-1} \end{aligned} \tag{10}$$

with the filtered residuals

$$r_k = (1 - H_k K_k)\, r_k^{k-1} \qquad R_k = (1 - H_k K_k)\, V_k . \tag{11}$$

K_k is sometimes called the gain matrix. The χ^2 contribution of the filtered point is then given by

$$\chi^2_{k,F} = r_k^T R_k^{-1} r_k . \tag{12}$$

The system state vector at the last filtered point always contains the full information from all points. If one needs the full state vector at every point of the trajectory, the new information has to be passed upstream with the smoother equations:

$$\begin{aligned} A_k &= C_k F_{k+1}^T (C_{k+1}^k)^{-1} \\ \tilde{x}_k^n &= \tilde{x}_k + A_k (\tilde{x}_{k+1}^n - \tilde{x}_{k+1}^k) \\ C_k^n &= C_k + A_k (C_{k+1}^n - C_{k+1}^k) A_k^T \\ r_k^n &= m_k - H_k \tilde{x}_k^n \\ R_k^n &= R_k - H_k A_k (C_{k+1}^n - C_{k+1}^k) A_k^T H_k^T . \end{aligned} \tag{13}$$

Thus, smoothing is also a recursive operation which proceeds step by step in the direction opposite to that of the filter. The quantities used in each step have been calculated in the preceding filter process. If process noise is taken into account, e.g., to model multiple

scattering, the smoothed trajectory may, in general, contain small kinks and thus reproduce more closely the real path of the particle.

In the equations above, F and H are just ordinary matrices if both transport and projection in measurement space are linear operations. In the case of non-linear systems, they have to be replaced by the corresponding functions and their derivatives:

$$F_k\tilde{x}_k \rightarrow f_k(\tilde{x}_k) \qquad H_k\tilde{x}_k \rightarrow h_k(\tilde{x}_k) \tag{14}$$

using for covariance matrix transformations

$$F_k \rightarrow \frac{\partial f_k}{\partial \tilde{x}_k} \qquad H_k \rightarrow \frac{\partial h_k}{\partial \tilde{x}_k}. \tag{15}$$

The dependence of f_k and h_k on the state vector estimate will in general require iteration until the trajectory converges such that all derivatives are calculated at their proper positions. We will continue to call $\partial f_k/\partial\tilde{x}_k$ the transport matrix and $\partial h_k/\partial\tilde{x}_k$ the projection matrix of our system.

The Kalman filter has also been found to be particularly suited for implementation in object-oriented programming languages [28].

2.5. Evaluation of performance

When it comes to quantifying the performance of methods in track pattern recognition, actual numbers will in general strongly depend on the definition of criteria, a fact that comparisons should take into account.

2.5.1. The reference set. Assessment of track finding efficiency requires, first, a definition of a reference set of tracks that an ideally performing algorithm should find. Normally, tracks will be provided by a Monte Carlo simulation, and the selection of reference tracks will depend on the physics motivation of the experiment. Low momentum particles arising from secondary interactions in the material are normally not within the physics scope but merely an obstacle and should be excluded. Particles travelling outside of the geometrical acceptance, for example, within the beam hole of a collider experiment, cannot be traced by the detector and should be disregarded as well. Also particles straddling the border of a detector and, e.g. traversing only a small number of tracking layers will often be regarded as outside of the design tracking volume. A typical convention may be to regard particles which traverse $\mathcal{O}(80\%)$ of the nominal tracking layers as constituents of the reference set.

The definition of the reference set can then be regarded as a definition of effective geometrical acceptance

$$\epsilon_{\text{geo}} = \frac{N_{\text{ref}}}{N_{\text{total}}} \tag{16}$$

with N denoting the number of particles of interest in the reference set and in total.

2.5.2. Track finding efficiency. The definition of the track finding efficiency requires a criterion which specifies whether a certain particle has been found by the algorithm or not. There are two rather different concepts:

Hit matching. This method analyses the simulated origin of each hit in the reconstructed track using the Monte Carlo truth information. If the qualified majority of hits, for example, at least 70%, originates from the same true particle, the track is said to reconstruct this particle. This method is stable in the limit of very high track densities, but it requires the Monte Carlo truth information to be mapped meticulously through the whole simulation.

Parameter matching. The reconstructed parameters of a track are compared with those of all true particles. If the parameter sets agree within certain limits (which should be motivated by the physics goals of the experiment), the corresponding track is said to reconstruct this particle. This method requires less functionality from the simulation chain, but there is a danger of accepting random coincidences between true particles and artefacts from the pattern recognition algorithm. In extreme cases, this can lead to the paradoxical impression that the track finding efficiency improves with increasing hit density.

The reconstruction efficiency is then defined as

$$\epsilon_{\text{reco}} = \frac{N_{\text{ref}}^{\text{reco}}}{N_{\text{ref}}} \tag{17}$$

where $N_{\text{ref}}^{\text{reco}}$ is the number of reference particles that are reconstructed by at least one track. It should be noted that this definition is such that a value of one cannot be exceeded, and multiple reconstructions of the same track will not increase the track finding efficiency. One should also control the abundance of non-reference tracks which are reconstructed ($N_{\text{non-ref}}^{\text{reco}}$): normally the relation

$$\frac{N_{\text{non-ref}}^{\text{reco}}}{N_{\text{total}} - N_{\text{ref}}} \ll \epsilon_{\text{reco}} \tag{18}$$

should hold, otherwise the reference criteria might be too strict.

2.5.3. Ghosts. Tracks produced by the pattern recognition algorithm that do not reconstruct any true particle within or without the reference set are called ghosts. The ghost rate can be defined as

$$\epsilon_{\text{ghost}} = \frac{N^{\text{ghost}}}{N_{\text{ref}}}. \tag{19}$$

Since the ghost rate may be dominated by a small subset of events with copious hit multiplicity, it is also informative to specify the mean number of ghosts per event.

2.5.4. Clones. The above definitions for efficiency and ghost rate are intentionally insensitive to multiple reconstructions of a particle. Such redundant reconstructions are sometimes called clones. For a given particle m with N_m^{reco} tracks reconstructing it, the number of clones is

$$N_m^{\text{clone}} = \begin{cases} N_m^{\text{reco}} - 1, & \text{if } N_m^{\text{reco}} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

and the clone rate becomes

$$\epsilon_{\text{clone}} = \frac{\sum_m N_m^{\text{clone}}}{N_{\text{ref}}}. \tag{21}$$

In practice, clones can usually be eliminated at the end of the reconstruction chain by means of a compatibility analysis [29].

2.5.5. Parameter resolution. The quality of reconstructed particle parameters and error estimates from reconstruction in a subdetector is essential for matching and propagation into another subsystem. For the whole detector, it determines the physical performance directly. The quality of the estimate of a track parameter X_i is reflected in the parameter residual

$$R(X_i) = X_i^{\text{rec}} - X_i^{\text{true}}. \tag{22}$$

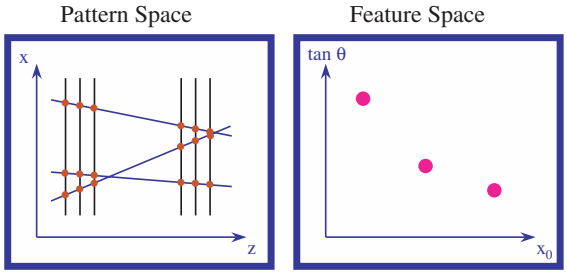

Figure 10. Schematic illustration of the pattern space (left) and feature space (right).

From the parameter residual distribution, one can then obtain the parameter estimate bias $\langle R(X_i)\rangle$, and the parameter resolution as a measure of its width. The estimate of the parameter covariance matrix can be used to define the normalized parameter residual

$$P(X_i) = \frac{X_i^{\text{rec}} - X_i^{\text{true}}}{\sqrt{C_{ii}}} \tag{23}$$

which is often called the pull of this parameter. Ideally, the pull should follow a Gaussian distribution with a mean value of zero and a standard deviation of one.

2.5.6. Interplay. Results for the individual performance estimators may very much depend on the definitions, so it is advisable to always judge several of the above quantities in combination. For example, the track finding efficiency should always be seen together with the ghost rate, since a less strict definition of the criterion if a track reconstructs a particle will lead to a higher track finding efficiency but also to a higher ghost rate. Also the parameter resolution will tell if the reconstruction criterion is correct, because in the case of an inadequately generous assignment, the parameter residuals are likely to show an increased width, or tails from improperly recognized tracks. When parameter matching is used, generous definition of the matching criteria will increase the track finding efficiency, but will also be reflected in the high clone rate.

Excessive tightening of the reference set criteria can also potentially ameliorate the visible track finding efficiency, but it will be at the cost of the effective acceptance, since the total yield of particles with a certain physical signature is proportional to the product

$$\epsilon_{\text{total}} = \epsilon_{\text{reco}} \cdot \epsilon_{\text{geo}} \tag{24}$$

always assuming that relation (18) holds.

3. Global methods of pattern recognition

The task of pattern recognition, in general, can be described by the illustration in figure 10. The physical properties of the particles that are subject to measurement are described by a set of parameters, such as point of origin, track direction or momentum. Each particle can, therefore, be represented by a point in the feature space spanned by these parameters. The signals the particle leaves in the electronic detectors are of a different kind; they are measured hit coordinates the nature of which is governed by the type of device. These coordinates are represented in the pattern space. While the conversion from feature to pattern space is done by nature, or by sophisticated simulation algorithms in the case of modelled events, the reverse procedure is the task of the combined pattern recognition and track fitting process.

Global methods assess the pattern recognition task by treating all detector hits in a similar way. The result should be independent of the starting point or the order in which hits are processed. This is unlike the local methods that will be discussed in section 4, which depend on suitable seeds for track candidates. Global methods aim to avoid any kind of seeding bias.

3.1. Template matching

The simplest method of pattern recognition can be applied if the number of possible patterns is finite and the complexity is limited enough to handle them all. In this case, for each possible pattern a template can be defined, for example, a set of drift chamber cells through which track candidates in a certain area will pass. Such a technique has been used for the Little Track Finder, which was part of the second trigger level of the ARGUS experiment [30], and which worked by comparing the hits in the drift cells of the axial layers to masks stored in random access memory. This method allowed for basic track finding in a two-dimensional (2D) parameter space, the track azimuth and the curvature in the R/ϕ projection, within 20 μs. The granularity of the ARGUS drift chamber was moderate, which limited the number of templates that had to be generated. The concept was later extended to the ARGUS vertex trigger [31], which used the hits of the micro-vertex detector [32] and generalized the algorithm to three dimensions and four parameters (track curvature being negligible), which allowed the measurement of the track origin in z to reject background interactions in the beam pipe. This algorithm required the definition of more than 245 000 masks, where a fivefold symmetry of the detector had already been exploited.

Template matching algorithms are mathematically so simple that they can be hard-wired as track roads, provided that the hit efficiency of each element is close to one. Remarkably, the computing time may be independent of the event complexity, since the number of templates to be checked is always the same. However, template matching does not scale very well when the problem requires high dimensionality or granularity. On the one hand, with increasing granularity, the number of templates already quickly exceeds the limits of feasibility where storing is concerned. Also, the number of computations increases significantly with a finer resolution of templates. Keeping the granularity low, on the other hand, means that dense situations cannot be resolved, and other methods have to be used to disentangle them.

An elegant solution to both problems is the tree-search algorithm, which uses templates of increasing structural resolution that are ordered in a hierarchy [33, 34]. In the first step, the hit structure is viewed at a very coarse resolution with a small set of templates (figure 11). For those templates that have 'fired', i.e. which match a structure prevalent in the event, a set of daughter templates with finer granularity is applied which are all compatible with the first matched template. This subdivision of templates is iterated until either a matching template on the finest level of granularity is reached, indicating that a good track candidate has been found, or a pattern matched at a certain resolution level cannot be resolved at the next level, in which case it is attributed to a random combination of hits.

The tree-search approach avoids the linear growth in the number of computations with increasing granularity that would develop in a purely sequential search; instead, the computing effort, at least for small occupancy, increases only logarithmically with the number of detector channels. The algorithm becomes even handier when storage of all possible templates can be avoided: in many cases, symmetries of the detector can be used to formulate rules relating to how the daughter templates can be derived from the parent at run-time, and how they are connected with the event data. The tree-search algorithm is used, for example, in the pattern recognition of the HERMES spectrometer, where the final detector resolution of 250 μm is reached in 14 steps [35]. Application of the tree-search ideally requires considerable simplicity

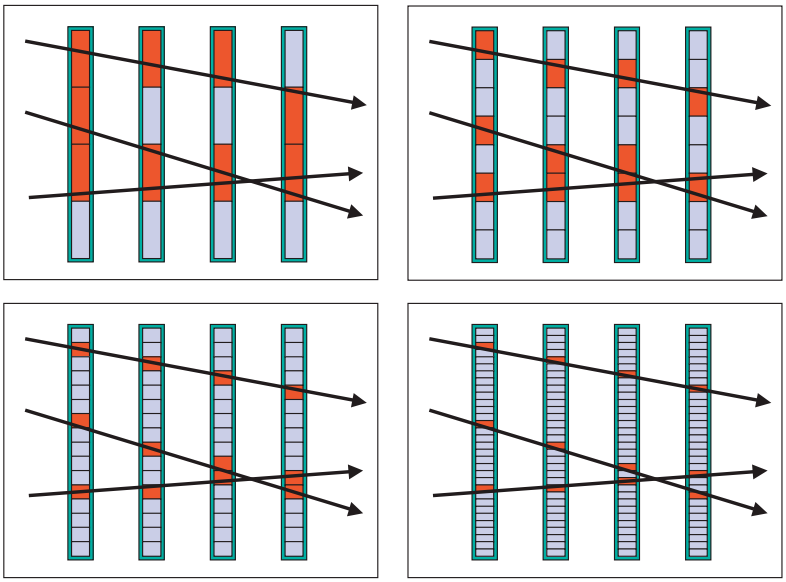

Figure 11. Schematic illustration of the tree-search algorithm: in several steps (in this case four), the track is matched with templates of increasing granularity and resolution. Each step descends into the next level of template hierarchy.

and symmetry in the detector design, and therefore cannot be easily used in many complex cases. In particular, inhomogeneous magnetic fields can complicate the application.

3.2. The fuzzy radon transform

In a very general sense, the observed hit density in the event can be described by a function $\rho(x)$, where x is a very general description of the measured set of hit quantities. In the absence of stochastic effects, the expected hit density in the pattern space can be described by an integral

$$\rho(x) = \int_P \rho_p(x) D(p)\,\mathrm{d}p \tag{25}$$

where $D(p)$ describes the prevalent population of the feature space, typically a sum of delta functions centred at the parameters of the particles, and $\rho_p(x)$ is the average response function in pattern space for a particle with parameters p, including all detector layout and resolution effects [36].

Pattern recognition can then be regarded as an inversion of the above integral from a stochastically distorted $\rho(x)$. The fuzzy radon transform of the function $\rho_p(x)$ is defined as

$$\tilde{D}(p) = \int_X \rho(x)\rho_p(x)\,\mathrm{d}x. \tag{26}$$

This transformation requires a precise knowledge of the response function, in particular the detector resolution. Track candidates are then identified by searching local maxima of the function $\tilde{D}(p)$.

This method will be illustrated using a simple example with a tracking system consisting of 10 equidistant layers in two dimensions without a magnetic field. Tracks are parametrized by an impact parameter x_0 and a track slope $t_x = \tan\theta_x$ as defined in section 2.3.1. As the measurement is 1D, each hit coordinate gives a linear warp-like constraint in the parameter

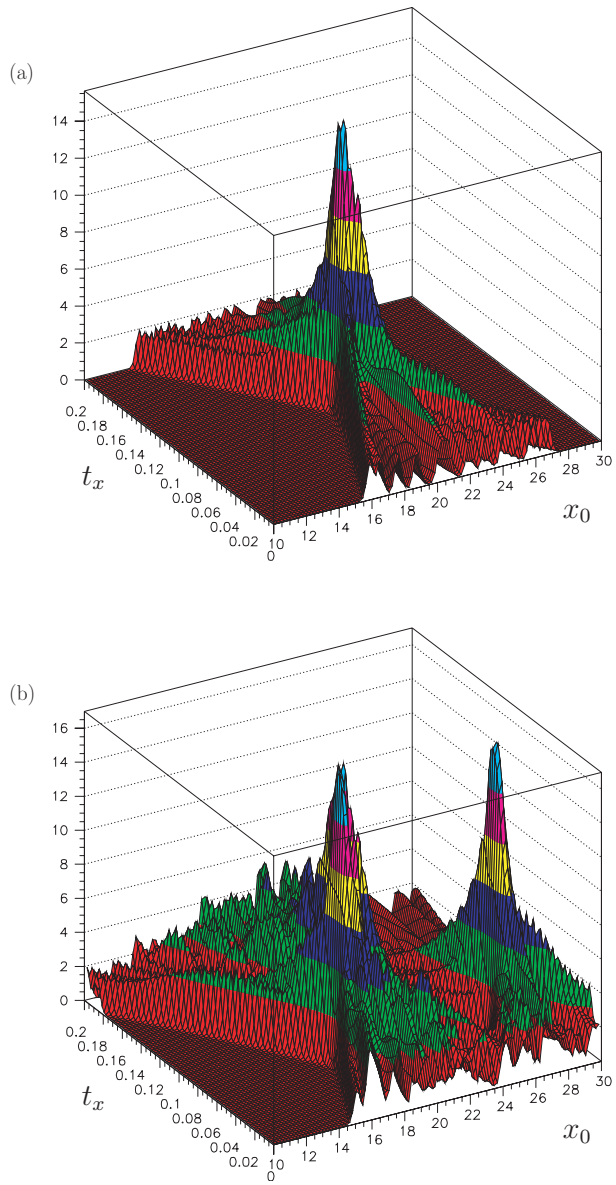

Figure 12. Fuzzy radon transform $\tilde{D}(x_0, t_x)$ of the hit signals of a single track (a), and in a scenario with three tracks (b), where x_0 and t_x are the track offset and slope.

plane, where the width of the warp reflects the effect of the detector resolution (figure 12(a)). For a fictitious situation with three superimposed tracks, the resulting fuzzy radon transform is shown in figure 12(b). The three peaks are very pronounced, but the development of additional local minima is already visible even in this clean situation.

In [36], this method has been explored for a cylindrical geometry, for the case of two very close tracks which only differ by a small difference in the curvature value (figure 13), with additional noise taken into account. Figure 14 shows the resulting radon transform $\tilde{D}(\kappa, \phi, \gamma)$ as a series of five images around the central values (γ stands for the z speed of the particle,

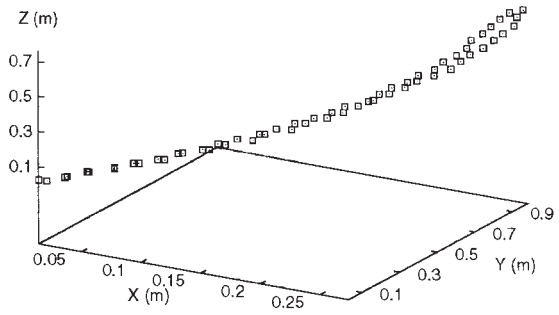

Figure 13. Two simulated tracks differing only by curvature (taken from [36]).

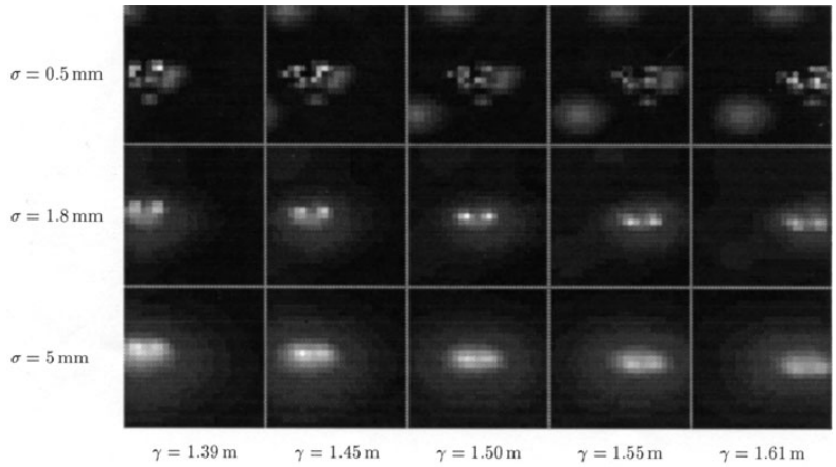

Figure 14. Fuzzy radon transform of the two tracks in figure 13 displayed in (κ, ϕ) space, with the third track parameter γ as described in the text (taken from [36]). The transform is shown for three values of the resolution parameter σ in $\rho_p(x)$, where the value in the middle row corresponds to the simulated resolution.

which is a measure of the dip angle tangent explained in section 2.3.2), where the resolution parameter σ has also been varied. The images show that the individual tracks can in fact be distinguished (centre image), but it is essential that the assumed resolution parameter matches the real one. It should be noted that automated recognition of the 'track signals' in such images would not be a trivial task, and that, for practical purposes, analysis of fuzzy Radon transforms in multi-dimensional parameter spaces are, in general, very demanding in terms of computing power.

Another generalization of the radon transform has been investigated in [37].

3.3. Histogramming

As seen in the previous section, the fuzzy radon transform allows us to take the precise detector resolution into account in an elegant manner. In cases where the effects of resolution can be neglected, the response function $\rho_p(x)$ only needs to describe the trajectory, and takes the shape of a delta-function whose argument vanishes for points on the trajectory. This special form of the radon transform is often called the Hough transform [39]. The Hough transform

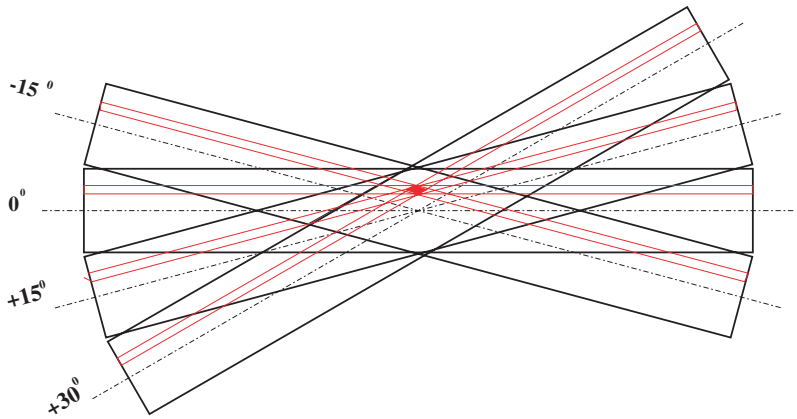

Figure 15. Illustration of wire orientations in the ZEUS STT. In this representation, the beam is oriented vertical to the page, displaced towards the bottom of the page (from [38]).

of each point-like hit in two dimensions becomes a line; more generally it defines a surface in the feature space. Completion of the pattern recognition task is thus converted into finding those points in feature space where many such lines or surfaces intersect, or at least approach each other closely in the shape of knots [39].

Histogramming can be regarded as a discrete implementation of the Hough transform. Hit information is converted to a constraint in a binned feature space, and the frequency of entries in a bin above a certain limit is indicative of a track candidate. However, in most tracking devices, a single measurement is not sufficient to constrain all track parameters. One solution is then to convert each measurement into a discretized curve or surface in parameter space, and to sample the contribution of all hits in corresponding accumulator cells. An example of such an implementation is shown for the straw-tube tracker (STT) of the ZEUS experiment [38]. This detector system is used as a forward tracker and consists of two superlayers with eight layers of straw tubes each. The straws are arranged in four different stereo views 0°, $\pm 15^\circ$ and 30°, as illustrated in figure 15. The 0° straws are oriented such that the point of closest approach to the beam line is in the middle of the straw. Taking the beam spot into account and neglecting the curvature of the segment within the confines of the straw tube tracker, each hit provides an arc-like constraint in the parameter space spanned by the polar angle θ and the azimuthal angle ϕ. This structure is displayed in the histogram from four views for a single track in figure 16. The hits from the 0° straws give a transform which has azimuthal symmetry, while the yields from the other views are slightly skewed with respect to the stereo angle. The parameters of the track are clearly indicated by the intersection of the four constraints. The resulting histogram is already much more complex in a sample with 10 simulated tracks, where combinatorial overlaps occur (figure 17).

Another popular way of avoiding the underconstrained case is to combine several hits to track segments before applying the Hough transform. For example, in a 2D pattern space without a magnetic field, two measured coordinates in the same projection from nearby hits in different detector layers give a straight track segment, which represents a point in the feature space. Histogramming all segment entries in the feature space should then reveal track candidates as local maxima. This procedure is often referred to as a local Hough transform [40].

In general, a price has to be paid for this artificial construction of a higher dimension of measurement, since random combinations of hits of different origin lead to ghost segments. The abundance of such contaminations depends strongly on the hit and particle density. A practical

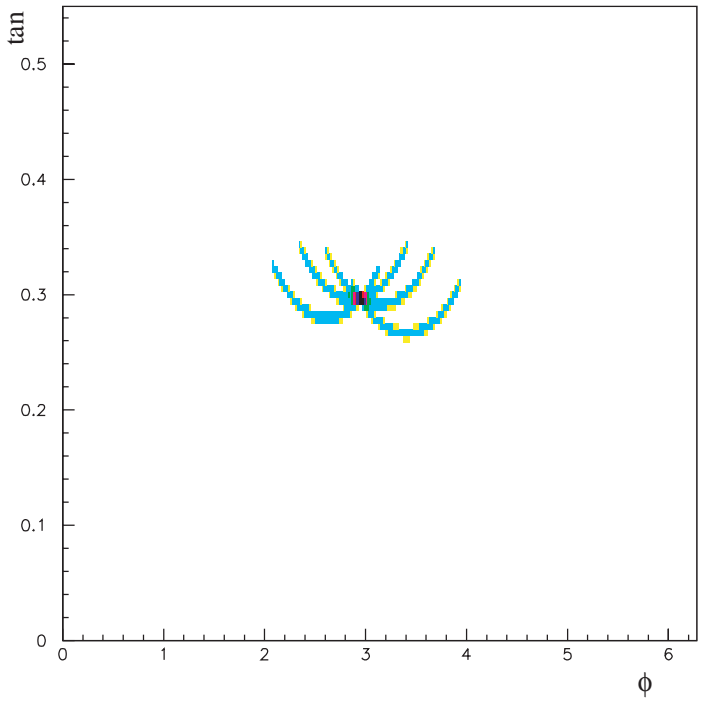

Figure 16. Hough transform of a single simulated track in the ZEUS STT (from [38]).

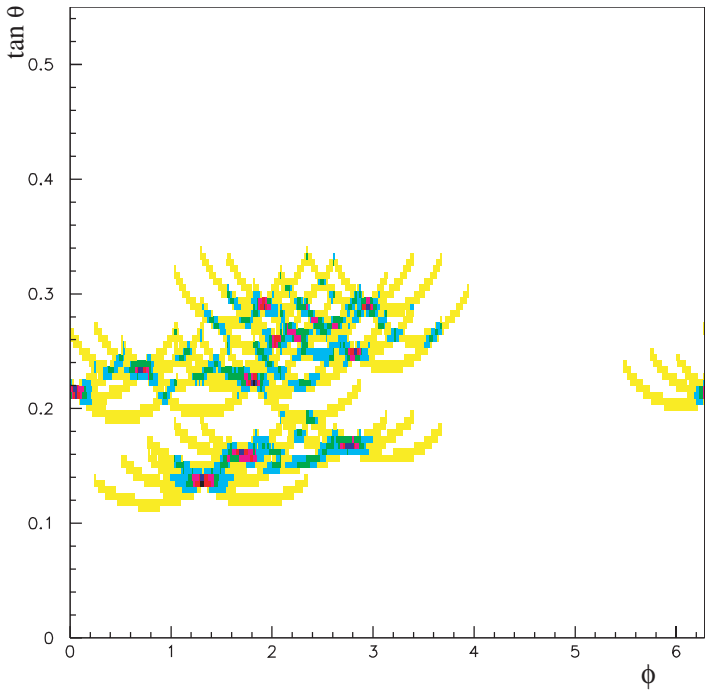

Figure 17. Hough transform of a set of simulated tracks in the ZEUS STT (from [38]).

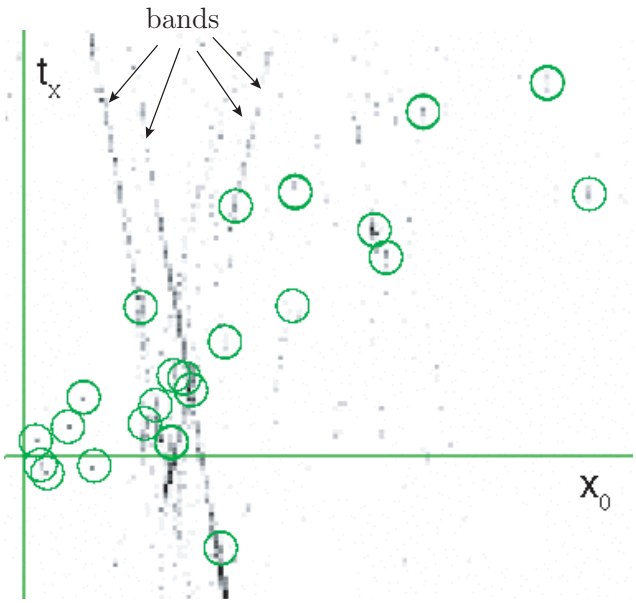

Figure 18. Local Hough transform in a simulated event with five interactions, in the feature space spanned by the impact parameter x_0 and track slope $t_x = \tan\theta_x$ (from [41]). The parameters of true particles are illustrated by circles. The colour intensity in each pixel corresponds to the count of segments falling into this square. While the histogram shows the expected enhancements at the true parameters of most simulated particles, it also displays artificial structures, indicated as bands in the plot that complicate the analysis.

example illustrating this problem is shown in figure 18 (taken from [41]). The geometry corresponds to the 'PC' part of the HERA-B spectrometer (see figure 8), which consists of four tracking superlayers, as indicated in figure 19(*a*), though in the latter the depiction has been simplified from six to three individual layers per superlayer. A simulated high-multiplicity event with five simultaneous pN interactions has been passed through a local Hough transform, a closeup of which is shown in figure 18. The genuine tracks as generated by the Monte Carlo process are indicated as circles in the feature space. While enhancements on the histogram are clearly seen at the track parameters of the true particles (indicated by circles), the histogram shows a significant number of bands that are caused by the interference of track patterns. Such interference occurs when several tracks cross the same superlayer of the tracking system within a close distance, as illustrated in figure 19(*b*) for four intersecting tracks: the proximity gives rise to a multitude of combinatorial segments, which have roughly the correct spatial information (x_{SL3}), but a wide range of deviating slopes shadowing the entries with the proper value. These segments enter the histogram with their spatial coordinate transformed to the reference plane relative to which all impact parameters are defined (in this case given by $z = z_{\mathrm{ref}}$) in the manner

$$x_0 = x_{\mathrm{SL3}} + (z_{\mathrm{ref}} - z_{\mathrm{SL3}}) \cdot \tan\theta_x . \tag{27}$$

The wide spread in the slope $\tan\theta_x$ results in a band in the parameter space, where the tilt of the band

$$\frac{\mathrm{d}\tan\theta_x}{\mathrm{d}x_0} = \frac{1}{z_{\mathrm{ref}} - z_{\mathrm{SL3}}} \tag{28}$$

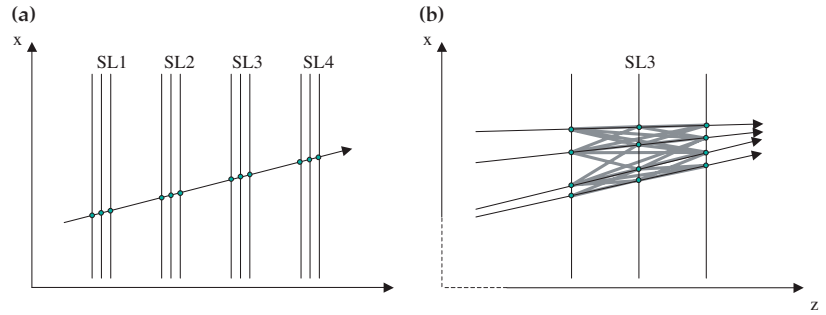

Figure 19. (*a*) Illustration of the model detector with four tracking superlayers discussed in the text, with the response of a single passing track. (*b*) Schematic illustration of track segments for a local Hough transform generated from four tracks intersecting in superlayer SL3, showing the abundance of ghost segments compared to the proper ones.

reflects the distance of the superlayer (at z_{SLi}) from the reference plane (at z_{ref}). It is, therefore, not surprising that in the given detector example with four superlayers, bands of four different slopes can occur.

Even in the absence of ghost segments from track overlap, the pattern of track signals in the discretized feature space will, in general, reflect the underlying layer structure of the tracking system. The local Hough transform is usually based on short segments, i.e. those composed of hits in subsequent or at least nearby layers, which has the advantage that the line topology of the track is exploited and the background from random hit combinations is still relatively small. However, due to the small leverage, the angular error can be sizeable, which may cause additional difficulty in identifying the track candidates in the Hough transform. Long segments spanning across many layers of the tracking system have the principal advantage of better angular resolution. However, a wide variety of hits have to be combined, so that the number of random combinations increases accordingly. The performance of different approaches has been analysed in detail in [42]. For an individual application, the optimal choice will depend on the relative importance of resolution and multiple scattering effects.

3.4. Neural network techniques

The human brain is particularly skilled in recognizing patterns. It is capable of analysing patterns in a global manner; it is self-organizing, adaptive and fault-tolerant. It is, therefore, not surprising that methods have been sought for, which aim at solving pattern recognition problems by means of artificial neural networks. Another intriguing aspect of the human brain is the massively parallel processing of information, which raises hopes that algorithms can be derived which can take full advantage of inherently parallel computing architectures. Because of the wide scope of this subject, this paper cannot give a full introduction to this field. A collection of classic papers that are reprinted is available in [43].

An artificial neuron manifests a simple processing unit, which evaluates a number of input signals and produces an output signal. A neural network consists of many neurons interacting with each other—the output signal of a neuron is fed into the input of many other neurons. While many classification problems can be attacked with simplified layouts, the feed-forward networks, track pattern recognition in general uses fully coupled topologies.

3.4.1. The Hopfield neuron. In the Hopfield model [44], each neuron, in general, interacts with every other neuron. All interactions are symmetric, and the state of each neuron, expressed

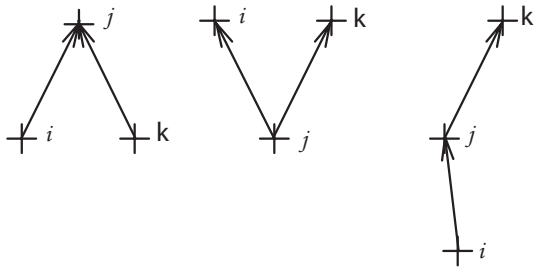

Figure 20. Three typical cases for adjacent track segments in the Denby–Peterson algorithm. The first two combinations correspond to incompatible segments; in the third case, both segments are likely to come from the same track (from [41]).

by its activation S_i, can only be either active (1) or inactive (0). The interaction is simulated by updating the state of a neuron according to the activations of all other neurons. The update rule in the Hopfield model sets the new state of a neuron to

$$S_i = \Theta\left(\sum_j (w_{ij} S_j - s_i)\right) \tag{29}$$

where the weights w_{ij} determine the strength of each interaction and s_i are the threshold values. The theta function $\Theta(\cdots)$, whose value is zero for negative arguments and one otherwise, is only the simplest example of an activation function, which relates the updated activation to the weighted sum of the other activations. It can be shown [44] that such interactions characterize a system with an energy function

$$E = -\frac{1}{2}\left(\sum_{ij} w_{ij} S_i S_j - 2\sum_i s_i S_i\right) \tag{30}$$

and that the interaction leads to a final state that corresponds to the minimum of the energy function [44,45].

3.4.2. The Denby–Peterson method. An adaptation of Hopfield networks to track finding has been developed by Denby [46] and Peterson [47]. The basic idea is to associate each possible connection between two hits with a neuron. Activation of such a neuron means that both hits are part of the same track. It is then essential to define an interaction such that in the global energy minimum only neurons corresponding to valid connections will be active. Interaction is only meaningful with neurons that have one hit in common. An approach to such an energy function is illustrated in figure 20 [41]: while in the first two cases the neurons (ij) and (jk) represent segments incompatible with the same track and, therefore, must have a repulsive interaction, the third case is much more track-like and should have an attractive interaction. This desired behaviour can be obtained by an energy function

$$E = -\frac{1}{2}\sum \delta_{jk} \frac{-\cos^m \theta_{ijl}}{d_{ij} + d_{jl}} S_{ij} S_{kl} + \frac{1}{2}\alpha\left(\sum_{l\neq j} S_{ij} S_{il} + \sum_{k\neq i} S_{ij} S_{kj}\right) + \frac{1}{2}\delta\left(\sum S_{kl} - N\right)^2 \tag{31}$$

where S_{ij} is the activation of the neuron associated with the segment (ij), i.e. the connection between hits i and j, and θ_{ijl} is the angle between the segments (ij) and (jl). The variables α and δ are Lagrange multipliers preceding terms that suppress unwanted combinations, such

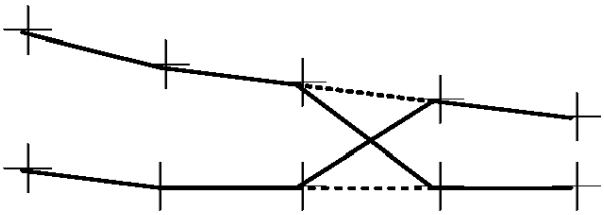

Figure 21. Wrong activations in the case of nearby tracks (from [41]).

as the first two cases in figure 20, and fix the number of active segments to the number of hits, N. Track finding is then reduced to finding the global minimum of this multivariate energy function. The interaction is simulated by recalculating the activity of each neuron with the update rule, which takes the activations of all other neurons into account.

It is remarkable that the Denby–Peterson method works without the actual knowledge of a track model—it favours a series of hits that can be connected by a line as straight as possible, but also allows small bending angles from one segment to the next. Thus, curved tracks can also be found, provided that a sufficient number of intermediate measurements exist, which split the track into a large number of almost collinear segments. The Denby–Peterson algorithm is, in particular, indifferent about the global shape of the track—a circle and a wavy track with the same local bending angles but alternating directions are of equal value.

One of the first explorations of the Denby–Peterson method has been performed on track coordinates measured by the ALEPH TPC [48]. The algorithm found tracks in hadronic Z^0 decays rather accurately, which may be at least partially attributed to three favourable circumstances: pattern recognition benefits considerably from the the 3D nature of the hits measured in the TPC, and equally from the clean event structure and the low occupancy. Moreover, the algorithm is applied such that the initialization activates only neurons that already correspond to plausible connections of hits. The authors of [48] have also investigated the behaviour of the method for events with much higher track numbers, simulated by piling up Monte Carlo events, and found that the total CPU time of the neural network algorithm is dominated by the initialization of the neurons, which indicates the degree of selection already involved at this stage.

The behaviour of the Denby–Peterson method under high track densities has been further investigated in [41] by applying it to a four superlayer geometry resembling the 'PC' part of the HERA-B tracker (see figure 8). These studies found that the classical Denby–Peterson method cannot be relied on to converge safely in cases of nearby parallel tracks. This behaviour is explained in figure 21; there is no possibility of resolving a cross-wise misassignment, since the system has reached a local energy minimum, and no additional segment can be attached because it would temporarily lead to an illicit branching of the track according to the rules illustrated in figure 20 and formulated in equation (31).

The situation can be improved, as shown in [41], by dropping the branching restriction and instead accounting for undesired angles in the cost function, by the replacement

$$\frac{-\cos^m \theta_{ijl}}{d_{ij} + d_{jl}} \rightarrow f(\cos^m \theta_{ij,kl}) \tag{32}$$

where the angle-dependent part is chosen such that only segments with angles close to 180° give a strong negative contribution, and by adding a term proportional to $(\delta - 1/d_{ij})$ for each neuron, which introduces a typical inverse segment length δ into the energy function, where the length of an individual segment d_{ij} is generalized such that the superlayer structure of the

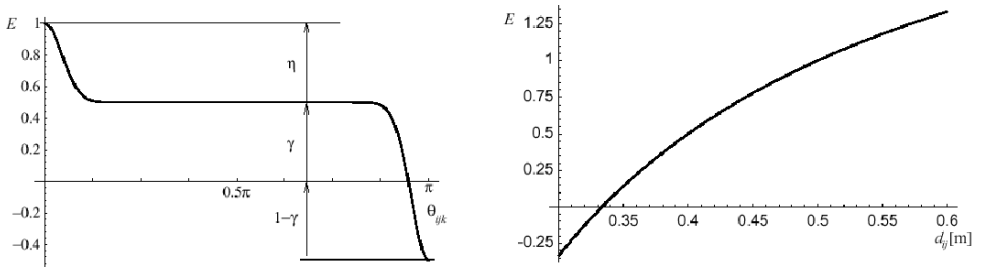

Figure 22. Modified energy function versus angle θ_{ijk} (left) and generalized segment length d_{ij} (right) as used in [41].

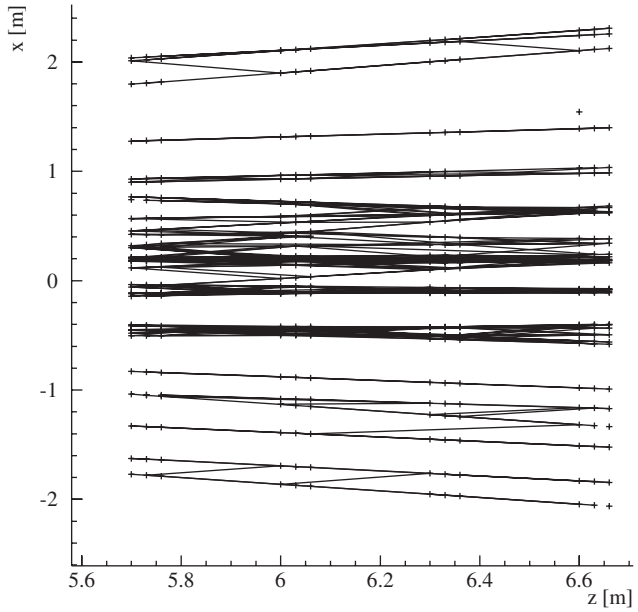

Figure 23. State of the network after one iteration [41]. Crosses denote the locations of the simulated hits.

tracker is taken into account. (The complete definitions are given in [41].) The energy as a function of segment angle and length is shown in figure 22.

The effect of this variation of the method is visible in figure 23, which shows the system after one iteration applied to an event with low track multiplicity. At this point, there are still branchings that would not be allowed in the classical Denby–Peterson approach, and which disappear under further iteration. With these modifications the algorithm obtains reasonable efficiency and ghost rate values [49, 41], as shown in figure 24.

Several properties of the Denby–Peterson algorithm limit its application at production scale in the general case. The fact that it does not take any explicit track model into account lets it ignore valuable information, which could otherwise help to resolve ambiguous situations. A straight track with random perturbations, e.g. is equivalent to a slightly curved track. Neither is there a way to explicitly take the resolution of the detector into account. The computing time per event increases with the third power of the track density, since the number of neurons that have to be generated is proportional to the number of hits squared, and the number of non-zero

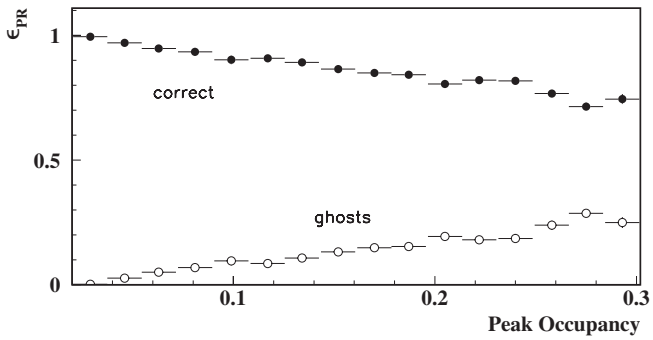

Figure 24. Efficiency and ghost rate for pattern recognition using the modified 2D Denby–Peterson algorithm on simulated events in a fixed target geometry (from [49]).

elements in the weight matrix increases with the number of neurons in the vicinity of the track. Perhaps the dominant shortcoming of the Denby–Peterson method is the fact that it does not have a direct extension for finding 3D tracks on the basis of single-coordinate measurements (see section 2.2.1), though it is in principle possible to circumvent this problem by first forming space points or segments out of the hits, provided that the ghost combinations are properly eliminated later. Such an approach has been successfully followed in [50], where a method resembling a discrete form of a Denby–Peterson net, referred to as a cellular automaton [51], was used to select optimal combinations of space points, complemented by a subsequent track following step.

3.4.3. Elastic arms and deformable templates. The above-mentioned limitations of the Denby–Peterson algorithm are overcome with the elastic arms algorithm [40, 52], which was introduced by Ohlsson *et al* in 1992. The basic idea can be described as follows: a set of M deformable templates is created, which correspond to valid parametrizations of tracks with parameters $\{t_1, \ldots, t_M\}$. The number M must be adjusted to the approximate number of tracks in the event. The algorithm should then move and deform these templates such that they fit the pattern given by the positions of N detector hits, which are represented by $\{\xi_1, \ldots, \xi_N\}$.

As in the Denby–Peterson case, the approach proceeds by formulation of an energy function, whose absolute minimum is at the set of parameters which solve the pattern recognition problem. This requires two elements: an activation-like quantity S_{ia} whose value is 1 if hit i is assigned to track a, and 0 otherwise, and a function $M_{ia}(\xi_i, t_a)$ describing a metric between track template and hit, typically the square of the spatial distance. The energy function can then be defined as

$$\tilde{E}(S, \xi, t) = \sum_{i=1}^{N} \sum_{a=1}^{M} S_{ia} M_{ia}(\xi_i, t_a). \tag{33}$$

To avoid trivial solutions, it is necessary to introduce the condition that each hit must be assigned to some template in the form

$$\sum_{a=1}^{M} S_{ia} = 1 \tag{34}$$

for each hit i. This requirement is called the Potts condition [53]. One immediate consequence of this condition is the necessity to introduce a special template to which noise hits can be assigned.

The main challenge is then to find the global minimum of the energy function. Since this function tends to be very spiky, as will be illustrated in more detail below, this problem is usually tackled by extending the energy function according to a stochastic model, which simulates a thermal motion in the system and smoothens out the spike structure. Search for the minimum, then, starts at high temperature, and the temperature is subsequently lowered. At zero temperature, the extended energy function becomes identical to the original one. This technique is called simulated annealing.

Instead of the temperature T, normally its inverse $\beta = 1/T$ is used. At finite temperature, the S_{ia} are replaced by their thermal mean values V_{ia}, which take continuous values and lead to a fuzzy hit-to-track assignment. They can be derived from the metric function as

$$V_{ia} = \frac{\mathrm{e}^{-\beta M_{ia}}}{\mathrm{e}^{-\beta\lambda} + \sum_{b=1}^{M} \mathrm{e}^{-\beta M_{ib}}} \tag{35}$$

where the index b in the sum in the denominator runs over all templates except for the noise template. V_{ia} is called the Potts factor. The temperature determines the range of influence for a hit: at zero temperature ($\beta \to \infty$), the hit is assigned only to the nearest template, with the corresponding V_{ia} equal to one. At a higher temperature, the degree of the assignment decreases smoothly with increasing distance. The noise parameter λ represents the symbolic noise template which, in the limit of zero temperature, takes over hits that are further than $\sqrt{\lambda}$ away from the nearest genuine template. It is, therefore, logical to set λ in correspondence to the detector resolution, typically as three or five standard deviations. The term $\mathrm{e}^{-\beta\lambda}$ accounts for assignments to the noise template. The Potts factor of the noise template is calculated using the equation

$$V_{i0} = 1 - \sum_{a \neq 0} V_{ia} \tag{36}$$

instead of equation (35), since the concept of a distance does not make sense here.

The only remaining steps necessary to solve the pattern recognition problem are

(a) to find a suitable initialization for the templates, and
(b) to find the absolute minimum of the energy function.

It turns out that both are non-trivial in practical applications. Before turning to realistic scenarios, it is very instructive to look at the shape of the energy function in a very trivial example (taken from [41]), which consists of a detector measuring only one spatial coordinate, namely x, and a track model consisting only of one parameter for each template. Two hits are considered with coordinates x_1 and x_2, and two templates with parameters x_a and x_b.

The energy as a function of the template parameters is shown in figure 25 at a high temperature (the hits being at coordinates $x_a = -1$ and $x_b = +1$). At this temperature, the templates perceive only a blurred image of the hit pattern. The global minimum is at the coordinates in the centre between the hits. When the temperature is lowered to a critical temperature T_c, a saddle point develops (figure 26), and the previous single minimum splits into two. The critical temperature is related to the coordinates as

$$T_\mathrm{c} = \frac{1}{\beta_\mathrm{c}} = \left(\frac{x_a - x_b}{2}\right). \tag{37}$$

At a very low temperature (figure 27), two minima have developed at positions corresponding to the two equally valid solutions, $x_a = x_1 \wedge x_b = x_2$ and $x_a = x_2 \wedge x_b = x_1$. The potential ridge at the line $x_a = x_b$ can be interpreted as a repulsive force between the templates [40].

The presence of the noise template parameter λ introduces further local minima into the energy function. An example is shown in figure 28 with three hits (with $x_c = 0.24$) and

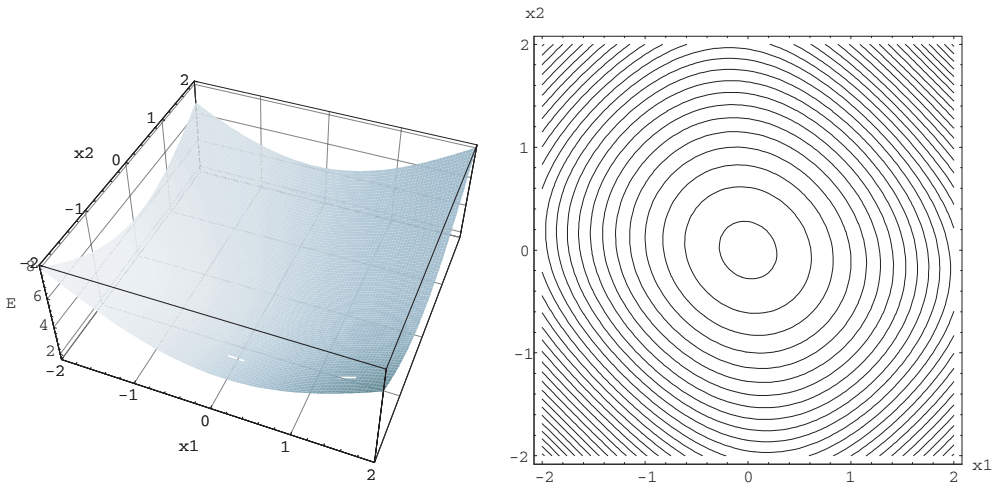

Figure 25. Representations of the energy function of a 1D detector with two hits, as a function of the parameters of two templates x_a and x_b at high temperature [41].

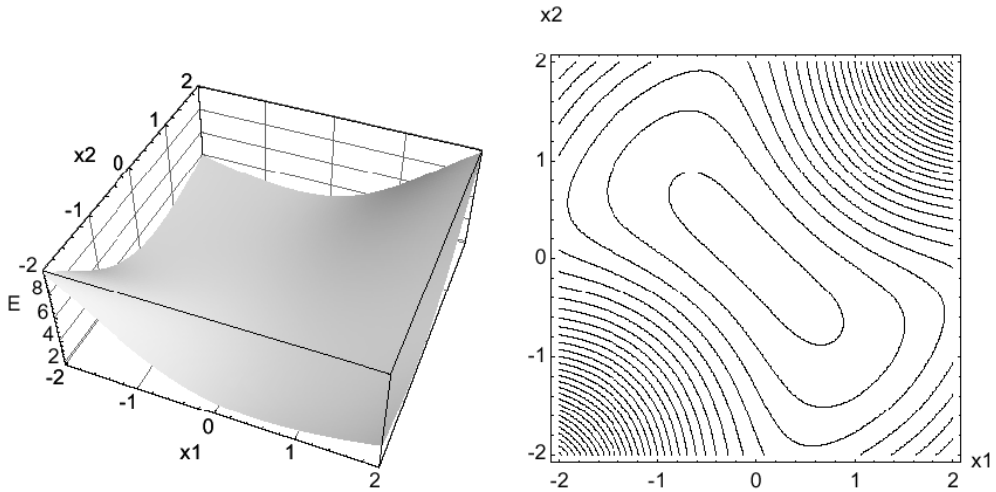

Figure 26. Energy function at critical temperature [41].

$\lambda = 0.4$. While the earlier solutions are still valid, additional minima appear that correspond to either one or two of the genuine hits being attributed to noise.

The complexity of the energy function for this very simple example is already staggering, and illustrates why initialization and convergence are serious issues.

In their initial study, Ohlsson *et al* [40] applied the method to hits from the DELPHI TPC. Reconstruction was restricted to tracks coming from a vertex spot common to all events, so that track candidates were described by only three parameters, which simplified the situation considerably. The initialization was obtained with a local Hough transform. The moderate hit density allowed the Hough transform to be performed first in the projection transverse to the magnetic field, searching for track candidates in the space of curvature and azimuth. For each

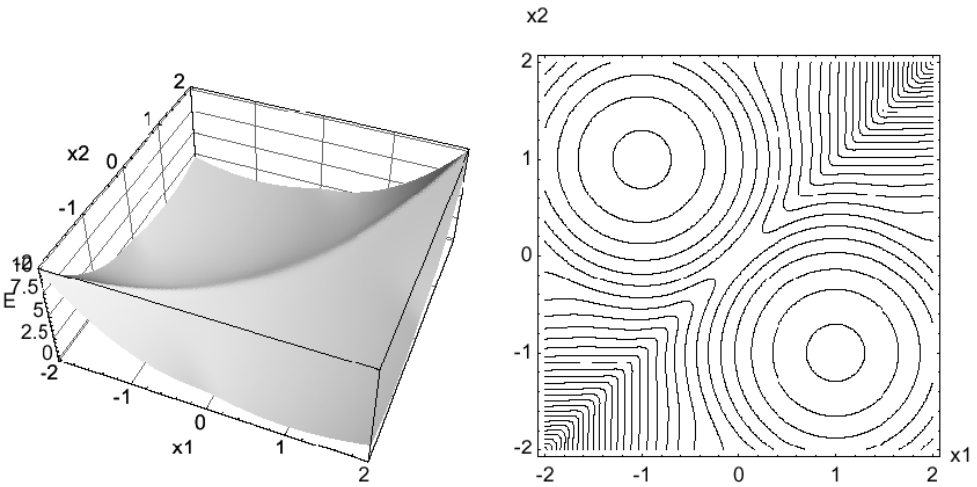

Figure 27. Energy function at a low temperature [41].

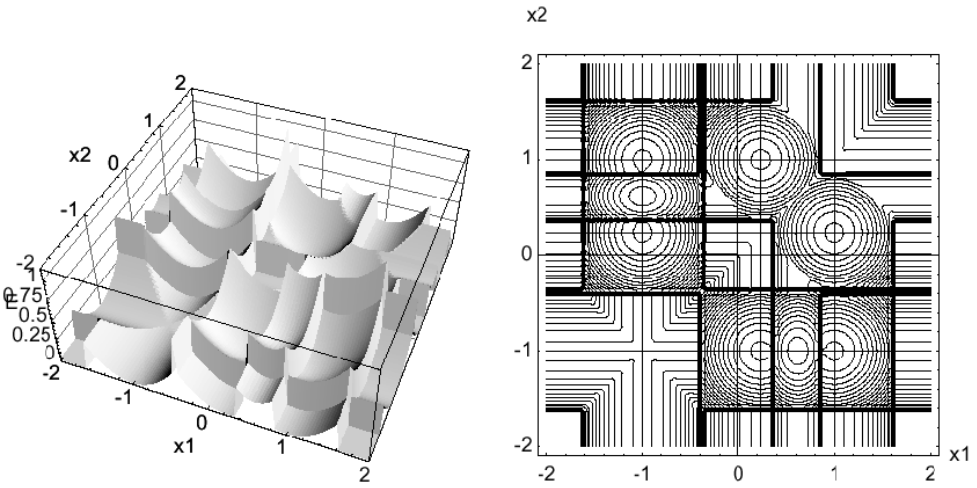

Figure 28. Energy function with three hits at low temperature, with $\lambda = 0.4$ [41].

candidate found as a narrow peak in this projection, all hits within a certain neighbourhood were used to calculate the longitudinal tilt angle, which was again histogrammed.

The elastic arms phase then used gradient descent to minimize the energy function at a given temperature. The temperature was lowered by 5% in each step. The Hough transform produced an abundance of templates. The excessive templates were either attracted to noise, or converged to tracks that already had templates associated with them; these had to be weeded out at the end. The result was found to be rather independent of algorithm parameters. The CPU time per event was dominated by the elastic arms step (1 min on a contemporary computer), in contrast to the Hough transform initialization (1 s).

Once more one has to note that pattern recognition in the TPC (here DELPHI's) benefits strongly from the clean event structure with a moderate track density, and the remarkable 3D measurement capabilities of the chamber. An interesting study targeted at much more dense

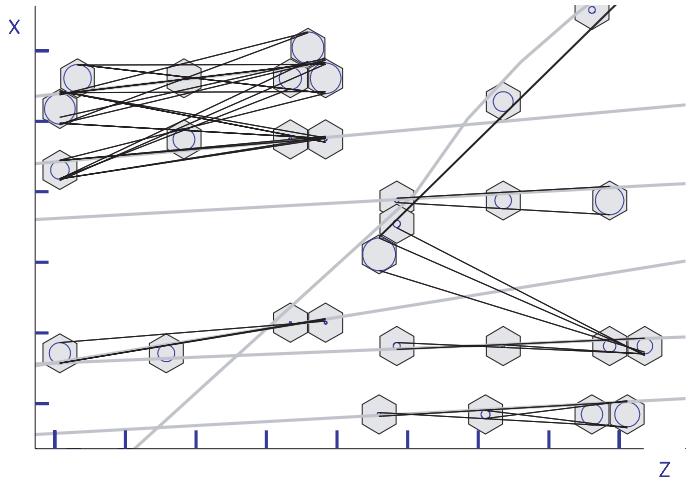

Figure 29. Illustration of segment initialization in the zx projection. The circles are drift distance isochrones of each hit with the drift cell indicated by a surrounding hexagon. The light grey lines are the simulated particles, the black straight lines connecting the hits are the segments produced to initialize the elastic arms algorithm [41].

events with 2D measurements was performed in 1995 [54]. The algorithm was applied to the barrel part of the transition radiation tracker (TRT) of the ATLAS detector, with 40 layers of straw drift-tubes with a diameter of 4 mm and a hit resolution of 150 μm. Since the required hit resolution could only be obtained using the drift-time measurement, the left–right ambiguity had to be resolved. This problem was addressed using the elegant method from [55], which introduces energy terms for both left–right assignments (in the nomenclature of equation (33))

$$\tilde{E}(S,\xi,t) = \sum_{i=1}^{N}\sum_{a=1}^{M} S_{ia}(s_{ia}^{+}M_{ia}^{+}(\xi_i,t_a) + s_{ia}^{-}M_{ia}^{-}(\xi_i,t_a)) \tag{38}$$

where the left–right assignment parameters $s_{ia}^{\pm}$, which satisfy the condition $s_{ia}^{+} + s_{ia}^{-} = 1$, introduce a repulsive interaction between the alternate left–right assignments, so that a track can only be assigned to one of the two ambiguities of a hit.

The initialization again used a local Hough transform. The minimization phase of the elastic arms step at a given temperature, however, did not rely on simple gradient descent, but used the Hessian matrix, i.e. the second derivative of the energy with respect to the parameters, in a multidimensional generalization of the Newton method. The efficiency was found to be 85% for fast tracks completely contained in the barrel TRT. The efficiency was practically identical to that of the Hough transform itself, indicating that the elastic arms part did not find any new tracks that had not been properly covered by the initialization. The main application of the elastic arms part was, therefore, to verify track candidates found by the Hough transform and resolve the hit associations.

The track finding capabilities of elastic arms have been further investigated in [41,56] with events passed through a full Geant simulation of the 'PC' area of the HERA-B spectrometer (see figure 8). Since the interpretation of the Hough transform turned out to be problematic in the fixed target geometry under study, a different approach was followed. Track candidates were initialized by searching hit triplets in the 0° projection in each of four superlayers (figure 29). All triplets with a straight-line-fit yielding $\chi^2 < 3.8$ were accepted, and then matched according to their track parameters. Combinations with triplet segments from all four superlayers were

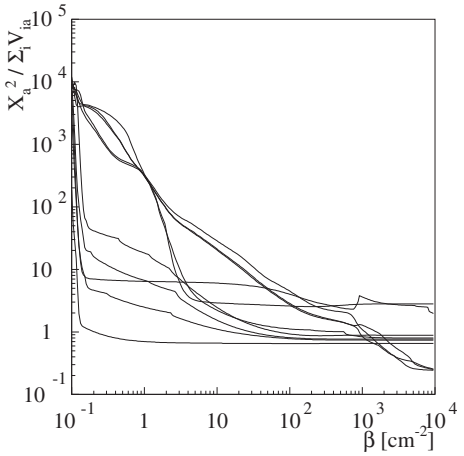

Figure 30. Development of χ^2 with increasing β (corresponding to decreasing temperature) with 10 muon tracks [41].

used to initialize the templates in the horizontal plane. The elastic arms algorithm was then used to perform the pattern recognition together with the stereo layers, arranging the tracks vertically. The proper operation of this method was shown with test events with 10 muon tracks, where the convergence of the tracks in the annealing from a temperature parameter of 0.1–10^4 cm^{-2} is illustrated in figure 30 by the decrease of the χ^2 per track. While the algorithm was actually performing the task of vertical pattern recognition after horizontal initialization, the computing time for the annealing with 10 tracks turned out to be already about 4000 s, and it increased at least with the second power of the number of templates. For this reason, dense events with 100 and more track candidates could not be seriously addressed with this method.

For this reason, a subsequent study [56] focused on the reduction of the processing effort. The first major step was the extension of the segment initialization to 3D. This was achieved by using the segments found from triplets in the xz projection to convert the information from the stereo layers to 3D coordinates: the segment in the projection defined a vertical plane in which the track candidate had to be contained (figure 31). Intersections of stereo wires with this plane lead to indirect measurements in the vertical coordinate y; the measurement equation

$$u[v] = x\cos\alpha - (\pm)y\sin\alpha \tag{39}$$

was inverted to

$$y = \pm\frac{x\cos\alpha - u[v]}{\sin\alpha} \tag{40}$$

and the triplet and segment finding proceeded with the stereo layers in a similar fashion. The stereo coordinates u and v took drift distance measurements into account, which improved the resolution but, also led to left–right ambiguities in the vertical segment finding.

The second crucial improvement concerned the minimization algorithm within each annealing step. The simplicity of the gradient descent method has made it highly popular for neural network applications, but as already observed in [54], it is not the most efficient method by far. One of its main drawbacks is the fact that its convergence slows down as it approaches the minimum where the surface of the energy function flattens out. On the other hand, large gradients as they can easily occur at lower temperatures (see figure 28) tend to

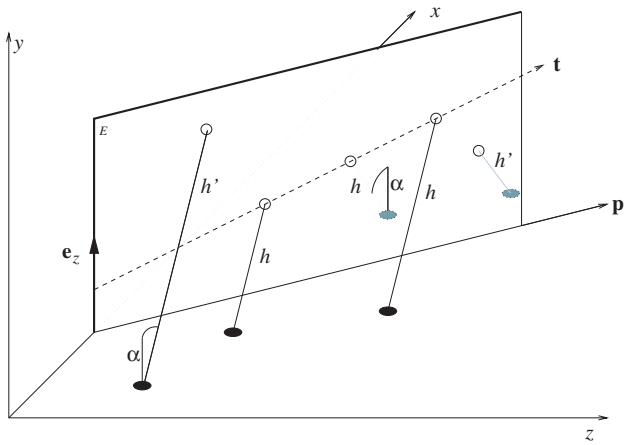

Figure 31. Scheme of converting information from stereo layers to vertical (y) coordinates, using the horizontal projection track candidate that is indicated by $\boldsymbol{p}$. The true track is indicated by a dotted arrow labelled $\boldsymbol{t}$. The ovals filled either in black or grey are coordinates measured under the stereo angle $\pm\alpha$ projected onto the xz plane. The lines labelled h are stereo hits stemming from the true track, these fall on the trajectory $\boldsymbol{t}$ when the y coordinate is inferred with equation (40). Other hits of different origin (labelled h') lead to background hits in the vertical plane (from [56]).

increase the step size drastically and throw the algorithm completely off the mark. These effects contribute largely to the high computing demands.

It is, therefore, promising to explore more efficient minimization techniques for high-dimensioned functions [56]. The Quickprop algorithm [57] parametrizes the dependence of the energy function on a template parameter $t_a^{(k)}$ (where a is the identifier of the template and k the index of the template parameter) in second order

$$E\left(t_a^{(k)}\right)_{\{\boldsymbol{t_a}\}} = c_0 + c_1 t_a^{(k)} + c_2 \left(t_a^{(k)}\right)^2 \tag{41}$$

and replaces the parameter in the iteration step $(i+1)$ with the value at the minimum of the parabola, which is calculated using the gradients of the two previous steps i and $(i-1)$:

$$\Delta t_{a,i+1}^{(k)} = \frac{-\partial E/\partial t_a^{(k)}|_i}{\partial E/\partial t_a^{(k)}|_i - \partial E/\partial t_a^{(k)}|_{i-1}} \Delta t_{a,i}^{(k)}. \tag{42}$$

Another more sophisticated minimization method, the RPROP algorithm [58], eliminates entirely the dependence of the step width of the gradient by using only its sign. Each component of the template parameter set has its own step width, which is reduced in each step if the sign of the partial derivative has not changed, and somewhat increased if the sign has changed, indicating a step across the minimum.

In applications to fully simulated events, the RPROP algorithm turned out to be 10 times faster than simple gradient descent. The Quickprop algorithm reduced the computing time by yet another factor of two, but failed to converge properly on about 10% of the tracks, so that the RPROP algorithm was finally chosen for further study [56].

The segment initialization achieved a track efficiency of 91% for single interactions, which dropped to 85% for five superimposed interactions in an event (table 1). The relative efficiency of the subsequent elastic arms phase was always better than 98%, indicating that hardly any of the good tracks that the initialization had found were lost. On the other hand, the elastic arms algorithm strongly reduced the rate of ghost tracks prevalent in the initialization. The CPU time consumption, determined on a HP9000/735 processor with a 125 MHz clock rate, was still

Table 1. Efficiency of segment initialization and elastic arms algorithm as compiled from [56], as a function of the number of superimposed interactions, N_{int}. In the elastic arms section of the table, efficiency, ghost rate and CPU time include the effects of segment initialization.

	Segment initialization			Elastic arms (incl. initialization)		
N_{int}	Efficiency (%)	Ghost rate (%)	CPU time (s)	Efficiency (%)	Ghost rate (%)	CPU time (s)
1	91	38	4	90	3.7	15
2	91	100	14	89	5.9	40
3	89	240	47	87	7.5	105
4	87	440	107	86	10	198
5	85	1100	234	83	13	371

relatively high, but with slightly more than 2 min for five simultaneous interactions already in a feasible range. With increasing track density the CPU fraction of the initialization increased steadily and exceeded that of the elastic arms part beyond three superimposed interactions.

The investigations underline that elastic arms can in principle be employed in an efficient manner, but require a very good initialization of the track candidates. This has led to the general perception that elastic arms should not be used for track finding from scratch, but should rather be seen as a tool to optimize assignment of hits to tracks, to resolve left–right or other ambiguities, or to detect and eliminate outlier hits. A similar philosophy is followed in [50]. A very interesting development in this context is the deterministic annealing filter (DAF) [59, 60], which extends the track fit with the Kalman filter with a fuzzy hit assignment and obtains a mathematical equivalent of the elastic arms procedure.

4. Local methods of pattern recognition

While global methods of pattern recognition have the common property to treat all hit information in an equal and unbiased way, simultaneous consideration of all hits can be very inefficient in terms of speed. In fact, many detector layouts provide sufficiently continuous measurements so that the sheer proximity of hits makes it already likely that they belong to the same track. This is one of the reasons why local methods of track pattern recognition, often called track following, are the workhorses of many reconstruction programmes in high energy physics.

Track following methods are essentially based on three elements:

- a parametric track model, which connects a particle trajectory with a set of track parameters and provides a method of transport, i.e. extrapolation along the trajectory;
- a method to generate track seeds, i.e. rudimentary initial track candidates formed by just a minimal set of hits which serve as a starting point for the track following procedure;
- a quality criterion, which allows distinguishing good track candidates from ghosts so that the latter can be discarded.

A related variant of track following is the propagation of a track candidate found in one part of the tracking system to another, collecting suitable hits on the way. In this case the initial track candidate takes the rôle of the seed.

4.1. Seeds

There are different possible philosophies on how seeds can be constructed. This is illustrated in figure 32, which shows schematically a tracking system with equidistant layers. Starting

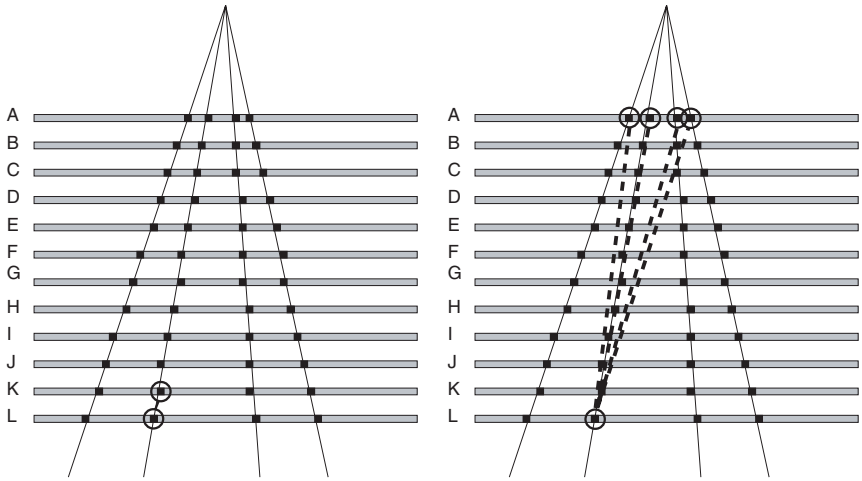

Figure 32. Seeding schemes with nearby (left) and distant layers (right).

from the last layer L, where the hit density is lowest, seeds can be obtained by combining the hit with suitable others in the neighbouring layer K (left side). This is the natural choice which exploits the local proximity of hits as a selection criterion. The angular precision of such a short segment is in general limited because of the small leverage, but the rate of fake seeds is relatively small, since most wrong combinations tend to obtain a steep slope that is incompatible with the relevant physical tracks and can be discarded immediately. A completely different alternative is to combine hits, for example, from the distant layers K and A to construct seeds. These seeds potentially have a much better precision in angle, but the number of choices to be considered is also much higher. The gain of precision can in fact be very limited if the material within the tracker introduces sizeable multiple scattering dilution. For the latter reasons, seed combinations from nearby layers are often preferred in practical applications.

Though the number of hits required for a seed is in general dictated by the dimensionality of the parameter space, additional hits can very efficiently decrease the ghost rate of the seeds. Figure 33 shows the construction of seeds consisting of three drift chamber hits each [61]. In this example without a magnetic field, only a minimum of two hits would be needed to define a seed, but the example shows that using hit triplets reduces the combinatorics considerably.

4.2. 2D versus 3D propagation

Many detector layouts allow track following in a projection. For example, drift chambers with many wires that are parallel and of the same length may allow separation of a pattern space that is measured in a plane orthogonal to the wires. This means that parameter propagation during the track following process is far less costly in terms of computations, and that the seeds can be constructed from only two measured hits in the case of a field-free area, or from three hits within a magnetic field. It should be noted that in the presence of a magnetic field, a 2D treatment is only possible if the field is oriented parallel to the wire, and homogeneous in the direction of the wire. An example of such an application is the pattern recognition in the ARGUS drift chamber (figure 34), where the seeds are constructed from three hits in the outer layers, and the track following proceeds towards the beam line [62].

However, pattern recognition in projections cannot avoid the fact that, at some point, 3D information must be inferred. This can be achieved by performing track finding independently

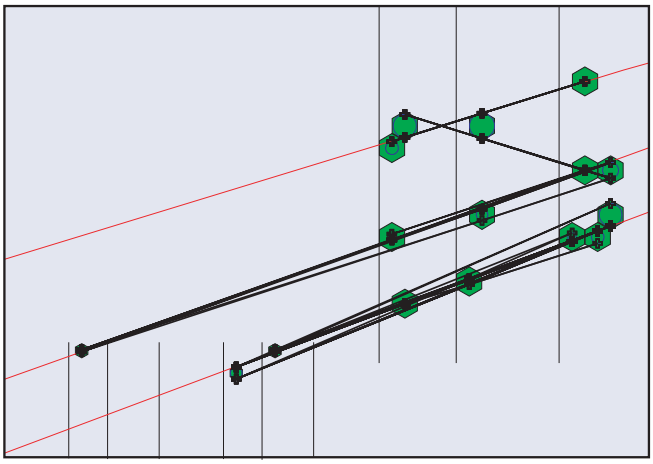

Figure 33. Creating seeds from drift chamber hit triplets. The style of displayed items is similar to figure 29. Crosses indicate the hit coordinates used to construct the triplets (from [61]).

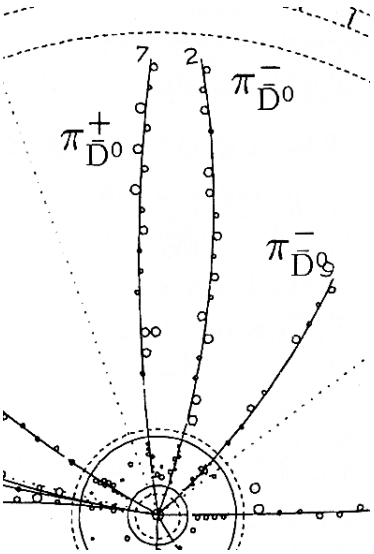

Figure 34. Close-up of the drift chamber area from the ARGUS event display in figure 1 [2]. The tracks are obtained by a track-following algorithm that proceeds from the outer towards the inner layers.

in all available projections, and then merging compatible projected track candidates into a 3D object. For an unbiased tracking, at least three independent views must exist (see section 2.2.3), and each view must possess enough hit information to find the track by itself with good efficiency. A typical symmetric arrangement consists of three views with 0°, 120° and 240° stereo angle, among which all layers are evenly distributed. This approach leads to virtually azimuth-independent track parameter resolutions.

A more economic alternative is a design with an asymmetric layer distribution which is less costly in terms of the number of channels but requires suitable pattern recognition algorithms.

It is possible to perform first the pattern recognition in the 0° projection, and then use the resulting track candidate to convert the measurements in the $+\alpha$ and $-\alpha$ layers into the vertical coordinate [62], as already illustrated in a different context in figure 31. The next step then proceeds with track finding in the vertical projection. In this case, only the 0° projection needs to be equipped with enough layers for a stand-alone track finding, while the two stereo views are combined and thus the number of layers per stereo view can be smaller. A reasonable scenario for this design comprises 50% of the layers oriented at 0°, 25% in the $+\alpha$ and 25% in the $-\alpha$ projection.

In the case of genuine 3D measurements, 3D seeds can be easily constructed from two hits in the field-free case, and from three hits in the case with a magnetic field, which normally will hardly lead to combinatorial problems. This is the situation in the barrel part of the CMS inner detector [63,64], where three layers of silicon pixel detectors of pixel size 150 μm will be used to initiate track seeds, or in TPCs. In the case of intrinsically 2D measurements, 3D seeding has the general disadvantage that the seeds will become rather complex, consisting of 4–5 measurements, and for a high particle density many false seeds will also be generated. Also left–right ambiguities have a strong impact here: a seed constructed from five drift chamber hits yields 32 ambiguous track parameter sets upon expansion of all possible left–right assignments. Once the seed is constructed, the track following step involves many extrapolations of the track parameters, which are more costly to implement with the full set of parameters, in particular if the covariance matrix is to be transported as well.

On the other hand, 3D propagation is easier to apply in the sense that the full coordinate information is always available, so that, e.g. the decision regarding whether the track candidate intersects a particular detector volume or not can be made unambiguously and multiple scattering effects can be accounted for with good precision. The issue of merging the different projections is also avoided.

4.3. Naïve track following

The naïve variant will be discussed here essentially to allow for comparison with the more sophisticated approaches. Starting from a seed, the trajectory is extrapolated to the detector part where the next hit is expected. If a suitable hit is found, it is appended to the track candidate. Where several hits are possible naïve track following selects the one closest to the extrapolated trajectory. This procedure is continued until the end of the tracking area is reached, or no further suitable hit can be found.

Naïve track following is relatively easy to apply to tracking scenarios with moderate track density and often leads to a reasonable computational effort since the number of hits to be considered is roughly proportional to both the number of layers and the number of tracks. The application to situations with large hit density soon reaches its limitations, since in dense environments, track following runs the risk of losing its trail whenever several possible continuations exist. The main complications can be summarized as follows:

(a) Some expected hits may be missing because of limited device efficiency, which will be called a track fault in the following. This also includes the case where the hit exists, but is out of expected coordinate bounds, for example, because of delta electrons created by the impact of the particle. In drift chambers with single hit readout, the drift time measurement of the followed track can be superseded by another particle passing the same cell closer to the signal wire.
(b) Wrong hits may be closer to the presumed trajectory than the proper hits and be picked up in their stead. This can happen easily just after the seeding phase when the precision of the track parameters is still limited, or when some false hits have already been accumulated.

A wrong hit may stem from another reconstructable track, from a non-reconstructable low-energy particle, or from detector noise.

(c) Left–right ambiguities in wire drift chambers double the number of choices. Especially in small drift cells, e.g. in straw tube trackers, wrong left–right assignments are to some degree unavoidable and need to be coped with.

These aspects can pose a particular problem if the track density is subject to strong variations, e.g. due to a fluctuating number of simultaneous interactions under LHC-like conditions.

4.4. Combinatorial track following

This variant is aware of possible ambiguities, and in each track following step, each continuation hit which is possible within a wide tolerance gives rise to a new branch of the procedure, so that in general a whole tree of track candidates emerges. The final selection of the best candidate must be done in a subsequent step, which may involve a full track fit on each candidate. This kind of method is potentially unbeatable in terms of track efficiency, but in general highly resource consuming and, therefore, only used in special cases with limited combinatorics.

4.5. Use of the Kalman filter

All track following approaches have to evaluate whether a certain hit is compatible with the presumed trajectory and thus suitable to be added to the track candidate. The suitability of a hit should be based on criteria that exploit all the knowledge based on those hits that have been accumulated so far. Not only the track parameters themselves, but their precision also needs to be known. The ideal tool in this situation is the progressive fit implemented by the Kalman filter, which has been discussed in section 2.4.2.

The Kalman filter prediction already provides an excellent criterion for hit selection. When a hit is considered to be appended to the track, first the predicted residual r_k^{k-1} from equation (9) can be used as a rough criterion. After passing a hit through the filter process (see equation (10)), the filtered χ^2 defined in equation (12) is an even more precise measure. In general, the decision power will increase when more and more hits are accumulated in the track candidate. Once the full track is available, the result of the Kalman smoother (equation (13)) can be used to detect and remove further outlier hits.

4.6. Arbitration

In practical applications of track following, means are required to reduce its dependence on the starting point, and to decrease its vulnerability against stochastic influences. This process is called arbitration. For example, it is mandatory not to depend on a single option of seeding tracks, which would lead to loss of a track if one of the seeding layers happens to be inefficient, but one will normally use several combinations of layers for seeding. Such redundancy increases the probability of obtaining a seed for a track even in the presence of device inefficiency. When an expected hit appears to be missing in a layer during propagation, it may be advisable not to discard the candidate immediately, but to proceed further until a fault limit is exceeded. In a case where more than one hit could present a suitable continuation for a track, one might not want to decide immediately for the closest hit but create branches into different candidates which are pursued independently. When a hit appears to be fine for a continuation, the algorithm should account for the possibility that this hit is wrong and the right hit has disappeared for some reason. However, naïvely applied, all these extensions

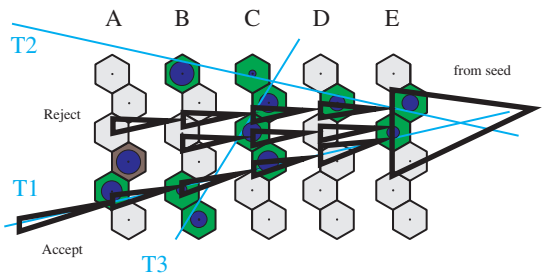

Figure 35. Schematic view of concurrent track evolution in a five-layered part of a tracking system with hexagonal drift cells, which is traversed by three particles, labelled T1, T2 and T3. The simulated drift time isochrones are indicated by circles. The propagation proceeds upstream from the right to the left and starts with a seed of hits from track T1 outside the picture (from [61]).

lead to either vast combinatorics, which will explode with increasing hit density, or suffer from *ad hoc* limitations. A method of overcoming these problems will be detailed in the following.

4.7. An example for arbitrated track following

This section discusses the concurrent track evolution algorithm as an example for an approach to track following with arbitration, which is described in detail in [61, 65].

4.7.1. Algorithm. The basic idea is to allow for concurrency of a certain number of track candidates at any time during the propagation of a certain seed, or even a set of seeds. These tracks are propagated in a synchronized manner from one sensitive tracking volume to the next. At each propagation step for each track candidate, branching into several paths is possible and will in general occur. Multiple branches appear when several continuation hits are consistent with the present knowledge of track parameters, or when more than one tracking volume is within reach. Also, the possibility that the expected hit is simply missing, e.g. because of device inefficiency, gives rise to a new branch. Thus, the procedure explores the available paths for all track candidates concurrently, which leads to a rapid creation of new track candidates. On the other hand, the number of track candidates should not grow beyond control. This is achieved by applying a quality selection on the whole set of concurrent track candidates after each round of propagation, using suitable estimators for the quality of a track. This leads to a favourable timing behaviour even for high multiplicity events. Concurrent track evolution can thus be regarded as a variant of deferred arbitration [66]. The actual propagation is based on the Kalman filter.

An illustration of this strategy is shown in figure 35 taken from [61], which shows a potentially ambiguous situation caused by two nearby tracks T1 and T2 plus a large angle track T3 in five layers of honeycomb drift chambers. For simplicity, it is assumed here that the algorithm discards track candidates with more than one missing hit (fault) in a row, and that the maximum number of concurrent candidates is three—in reality, higher limits may be used. It is also assumed that a seed of hits from track T1 has been formed on the right-hand side outside the figure. The propagation proceeds upstream from right to left. The illustration shows how three parallel candidates arise from different left–right assignments to the two drift chamber hits in layer E, which are propagated through layers D and C—including the tolerance of a fault on track T1 in layer D. In layers B and A, the false paths are discarded

Table 2. Table of parameters used in the implementation in [61].

Parameter	Value	Parameter	Value
$N^{\min}_{\text{Hits}}(x)$	9	$N^{\max}_{\text{Faults}}(x)$	2
$N^{\min}_{\text{Hits}}(y)$	9	$N^{\max}_{\text{Faults}}(y)$	2
$\delta\chi^2_{\max}(x)$	8	$\mathcal{R}_{\max}$	5
$\delta\chi^2_{\max}(y)$	16	w_{χ^2}	0.1
$\delta q_{\min}$	-1		

because of accumulating too many faults, and the proper reconstruction of track T1 is retained. Track T2 should then be found later with a different seed, while track T3 is likely to be non-reconstructable.

Track following in the naïve sense will always accept the hit with the smallest χ^2 contribution, which is possibly a good solution when the hit density is small. In the presence of multiple scattering and high hit densities, a wrong hit will frequently have a smaller χ^2 contribution than the proper one, or replace a proper hit which is missing due to detector inefficiency, or shadowed by another track passing the same cell. On the other hand, full evaluation of all possible hit combinations would exceed all bounds of computing resources when applied to dense events. Thus, the concurrent track evolution strategy combines the virtues of track following and combinatorial approaches. As will be shown below, the optimization in each evolution step using a quality estimator provides an elegant means to deal with the main problems in high occupancy track propagation.

4.7.2. Parameters. The algorithm is controlled by parameters which determine the selection of hits for the propagation of candidates, and for optimization of concurrent candidates on each level. The parameter $\delta_u^{\max}$ is the range around the predicted coordinate in the next considered tracking layer, in which continuation hits are searched. The parameter $\delta\chi^2_{\max}$ stands for the maximum tolerable filtered χ^2 increment according to equation (12). Missing hits (faults) are in general tolerated but only a certain number of subsequent faults ($N^{\max}_{\text{Faults}}$) are accepted. The pruning of track candidates after each evolution step is then regulated with absolute and relative cuts. The quality of a track candidate can be estimated with a function of the form

$$Q = f(N_{\text{Steps}}, N_{\text{Faults}}, \chi_i^2, \ldots) \tag{43}$$

where N_{Steps} is the number of evolution steps passed so far, and χ_i^2 stands for the contribution of the accumulated hit i to the total χ^2. If necessary, a bias from the track parameters could also be introduced here, which suppresses, e.g. tracks that are very steep or have very low momentum. A convenient simple quality estimator is

$$Q = N_{\text{Steps}} - N_{\text{Faults}} - w_{\chi^2} \cdot \sum_i \chi_i^2 \tag{44}$$

which applies a certain malus (in this case 1) for each missing hit, which is equivalent to an ill-matching hit with a χ^2 contribution of $1/w_{\chi^2}$ (in the configuration of table 2 equal to 10). Furthermore, cuts are applied relative to the best candidate currently in the set: candidates whose quality differs from the best candidate by more than $\delta q_{\min}$ are discarded. Finally, all concurrent track candidates are ranked in decreasing order of quality, and only the first $\mathcal{R}_{\max}$ candidates in rank are retained. If propagation cannot be continued though the end of the tracking system is not reached, this may have a natural reason, e.g. the particle may have been stopped or may have decayed in flight. In such cases, the best remaining track candidate on the last level is kept if it comprises at least a certain minimum number of hits, $N^{\min}_{\text{Hits}}$.

4.8. Track following and impact of detector design parameters

The practical behaviour of such an algorithm, as it has been developed for the HERA-B spectrometer has been studied in [61], including an investigation of the impact of detector design and performance on the pattern recognition capability. As the experiment has never routinely taken physics data at the high design interaction rate of 40 MHz, the results have been obtained from a full Geant simulation with, on average, five superimposed pN interactions, one of them containing beauty hadrons. As seen in figure 8, the inner part of the HERA-B main spectrometer acceptance within about 25 cm radius from the beam line is covered by MSGC, while the outer part is instrumented with Honeycomb drift chambers [13–15]. The pattern tracker consists of four superlayers outside the magnetic field, which consist of six individual layers each (the area marked 'PC' in figure 8), except for the inner part of the two middle superlayers that have only four layers each. Half of the layers measure a horizontal coordinate (0° orientation), the other half are arranged at a stereo angle of ± 100 mrad. The seeds were produced from hit triplets in the hindmost two superlayers for upstream, and in the foremost two superlayers for downstream propagation (figure 33). Track finding was performed first in the 0° projection, then continued in the combined stereo layers, where the vertical coordinates were determined using the horizontal projection of the track candidate, with the method explained in section 3.4.3 (see equation (40) and figure 31).

The algorithm parameters used are summarized in table 2. The parameters allow for a delicate adjustment of balance between the extremes of naïve track following ($\mathcal{R}_{\max} = 1$), where always the apparently best path is followed, and combinatorial track following ($\mathcal{R}_{\max} = \infty$), which retains all paths. The detailed simulation allowed us to study some principal effects of tracking system properties on pattern recognition parameters, which will be shown in the following.

4.8.1. Influence of detector efficiency. Figure 36 shows how the hit efficiency of the detector devices affects the pattern recognition performance on tracks emerging from B decays. Above $\epsilon_{\text{Hits}} = 95\%$, the hit inefficiency is well compensated by the algorithm (operating with $N_{\text{Faults}} = 2$), resulting in an excellent track finding performance. A smaller hit efficiency leads to a sizeable loss in the fraction of detected particles.

4.8.2. Effect of detector resolution. The influence of the spatial resolution is shown in figure 37. The simulated resolutions of the outer and inner tracking systems were varied independently. It is interesting to see that the efficiency degrades only slowly with the resolution being increased up to 1 mm. The slight drop in efficiency at 100 μm in figure 37(a) is an artefact due to numerical approximations. Both figures indicate that the effect of resolution on track finding efficiency should not be overrated. The effect on the ghost rate is much stronger; the plots underline that a good resolution helps considerably in suppressing fake reconstructions.

4.8.3. Influence of double track separation. The simulation of the inner tracker devices allowed varying of the double track resolution, i.e. the distance down to which nearby tracks can be resolved as individual hits in a device. In MSGC as they are used by HERA-B, the double track separation distance is in general larger than the resolution, since it depends on the cluster sizes. As visible in figure 38, the efficiency drops significantly with double track resolutions worse than 800 μm.

4.8.4. Execution speed. As already seen in section 3.4.3, the CPU time consumption is an essential aspect for the selection of a pattern recognition algorithm. The concurrent track

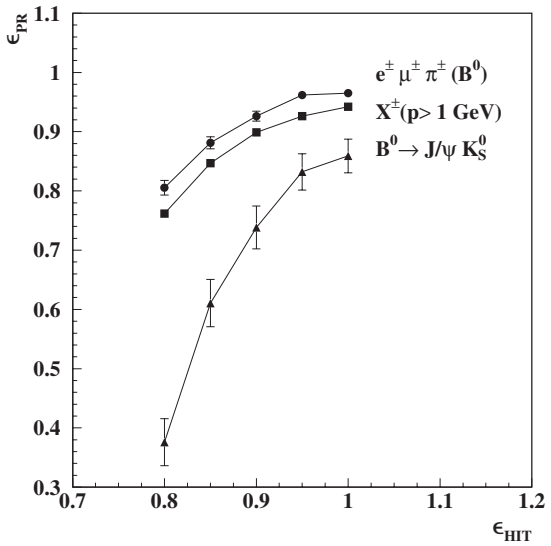

Figure 36. Pattern recognition efficiency for different values of the hit efficiency on simulated events consisting of one pN interaction with a B^0 meson with the decay chain $B^0 \rightarrow J/\psi K_S^0 \rightarrow \ell^+\ell^-\pi^+\pi^-$, where $\ell^+\ell^-$ can be a pair of muons or electrons, superimposed with, on average, four unbiased inelastic interactions. The filled squares (■) show the track finding efficiency for charged particles with momentum above 1 GeV; the filled circles (●) are for particles from the B decay. The triangles (▲) indicate the combined efficiency of all four B decay particles [61].

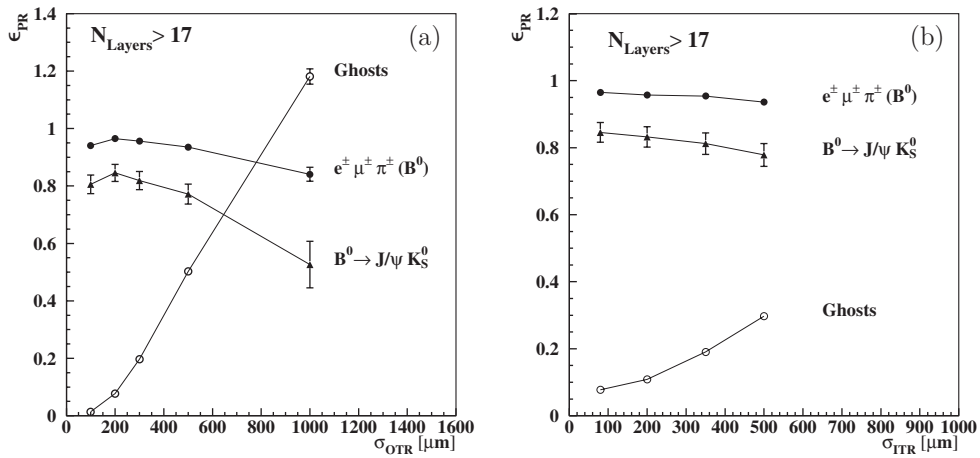

Figure 37. Pattern recognition efficiency for different outer (*a*) and inner tracker resolutions (*b*), for particles from the $B^0 \rightarrow J/\psi K_S^0$ decay mode as detailed in the caption of figure 36. Only tracks passing at least 17 out of 24 possible tracking layers were considered. Also the ghost rate is displayed (○) [61].

evolution algorithm was tested on the same geometry and event type as the elastic arms algorithm implementation (see table 1), and required on average 4 s per event with four superimposed inelastic interactions, compared to several minutes for the elastic arms method on the same type of processor. Also, the behaviour with increasing track density is important, since steep increases with a sizeable power of the track multiplicity, as they may arise from

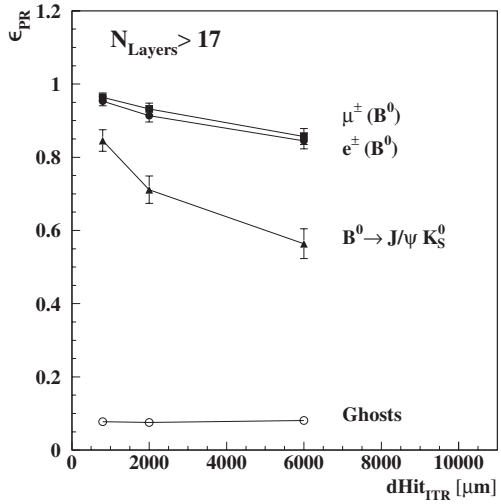

Figure 38. Pattern recognition efficiency for different double hit resolutions of the inner tracker for particles from the $B^0 \rightarrow J/\psi K_S^0$ decay mode as detailed in the caption of figure 36. The efficiency for muons (■) and electrons (●) is shown separately. The ghost rate is also shown (○). Only tracks passing at least 17 out of 24 possible tracking layers were considered [61].

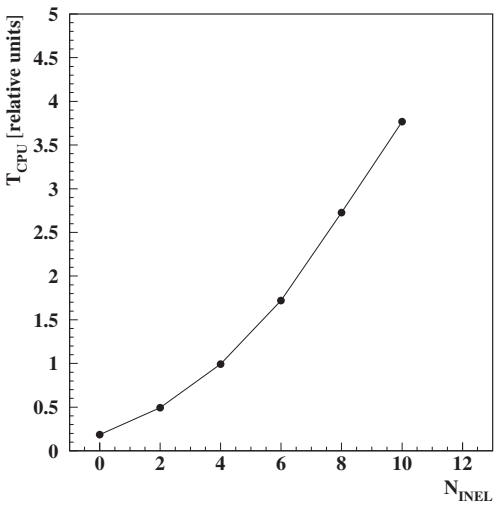

Figure 39. Mean computing time per event as a function of the number of inelastic interactions superimposed on one $b\bar{b}$ interaction [61], normalized to the value at $N_{\mathrm{INEL}} = 4$.

combinatorics, can have a very negative impact when a reconstruction program is used at production scale. Figure 39 shows the average computing time per event normalized to that for the nominal four superimposed inelastic interactions. At high interaction multiplicity, the computing time per event settled rather gracefully on a roughly linear dependence, indicating a constant amount of time per track at an acceptable loss of efficiency, which can be considered to be the desired behaviour. With the speed shown above, the algorithm is fast enough to be used in quasi-online reconstruction [67].

4.9. Track propagation in a magnetic field

In general, the above track following strategy can also be applied within a magnetic field. The main difference is that the transport function in equation (8) becomes non-linear, and the transport matrix becomes a local derivative as shown in equation (15). If the field is homogeneous, or if inhomogeneity can at least be neglected within typical transport distances, the transport function and matrix can usually be expressed analytically.

In many cases, however, the field is neither homogeneous nor can it be described by an analytic expression; instead, it is parametrized in terms of a field map, which has been measured with Hall probes, or computed by means of a field simulation program. In this case, numerical methods have to be used to derive the transport function. One very suitable method is the Runge–Kutta procedure [68], which integrates the equations of motion by expanding the trajectory up to a certain order and sampling the field at a series of intermediate points, which are chosen and weighted such that all powers of the errors below a certain order cancel. Even this procedure meets considerable challenges when the field varies strongly and a very high precision, matching the detector resolution, is warranted. In this situation, an embedded Runge–Kutta method with adaptive step size can help: the next highest order of Runge–Kutta is compared with the preceding one and the difference serves as an error estimate, which is then used to adjust the step size.

Application of the Kalman filter does not only require a transport function for the track parameters, but the derivative matrix of the new parameters with respect to the old is also needed (see equation (15)). Calculation of this derivative matrix can be efficiently performed within the same Runge–Kutta framework that is used for the parameter transport itself [69].

An extension of the concurrent track evolution algorithm for track following in the magnetic field has been developed and tested on the HERA-B geometry in [65]. Track segments found in the field-free part of the spectrometer were followed upstream through the inhomogeneous field of the magnet tracker. Figure 40 shows an event display with simulated tracks including a B decay reconstructed with this method. The algorithm achieved a high track propagation efficiency in spite of the large track density.

5. Fitting of particle trajectories

After pattern recognition has done its work, the detector hits are separated into sets, each of which, ideally, contains manifestations of one specific particle. It is then the task of the track fit to evaluate the track parameters and thus the kinematical properties of the particle with optimal precision. Even if the pattern recognition itself is already providing track parameters and covariance matrices to some degree, obtained, for example, by means of the Kalman filter, it will in general be left to a final track fit to take all necessary effects into account which are often neglected at the track finding stage because they are costly to apply under the full combinatorics of pattern recognition.

5.1. Random perturbations

In the easiest case, track parameters could be derived from the measurements by applying the least squares fit formulae from equations (4) and (5) in section 2.4.1. In realistic applications, the problem is usually more involved because of the way the trajectory of the particle is influenced by random perturbations that dilute the information content of the measurements, most commonly multiple scattering and ionization or radiative energy loss. Their influence is schematically displayed in figure 41. One can interpret the diagram in such a way that, from

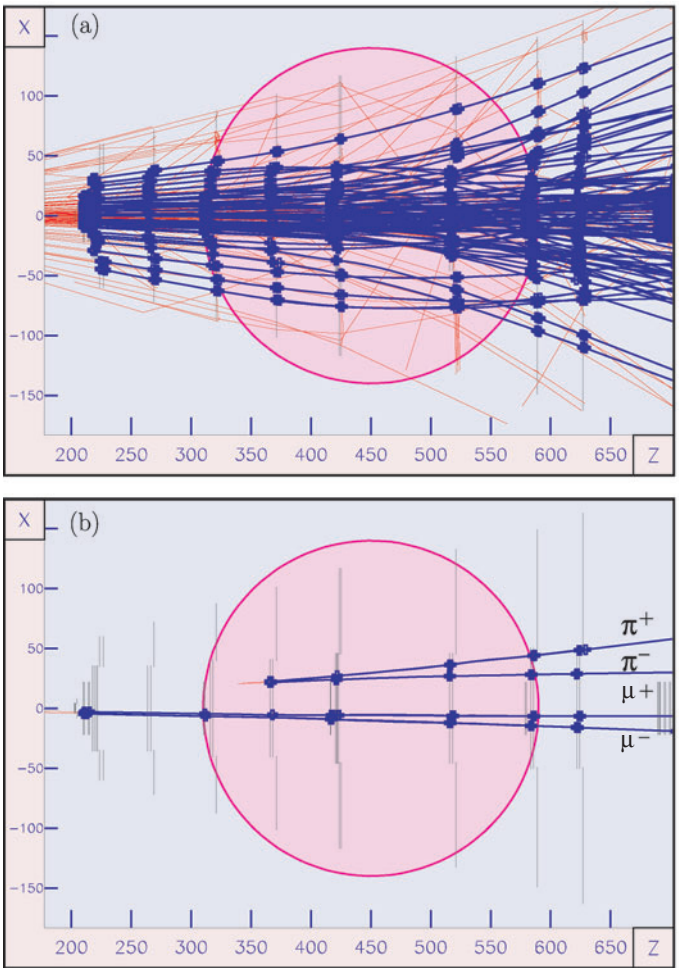

Figure 40. (*a*) Display of a simulated event with one interaction containing the golden B decay and six superimposed inelastic interactions, focused on the magnet area, where the pole shoe of the magnet is indicated by the large circle [65]. Both the Monte Carlo tracks (light grey) and the reconstructed tracks (thick dark lines) are shown (reconstructed hit points denoted by crosses). (*b*) Same event, with the display restricted to particles from the golden B decay.

step to step, the measurements, labelled on the right-hand side, improve the degree of amount of information about the kinematical properties of the particle, while the perturbations labelled on the left side reduce it.

5.2. Treatment of multiple scattering

Multiple scattering occurs through the elastic scattering of charged particles in the Coulomb field of the nuclei in the detector material. Since the nuclei are usually much heavier than the traversing particles, the absolute momentum of the latter remains unaffected, while the direction is changed. If the longitudinal extension of the traversed material block can be neglected (this is normally referred to as thin scatterer approximation), only track parameters related to particle direction are affected directly, for example, the track slopes $t_x = \tan\theta_x$ and

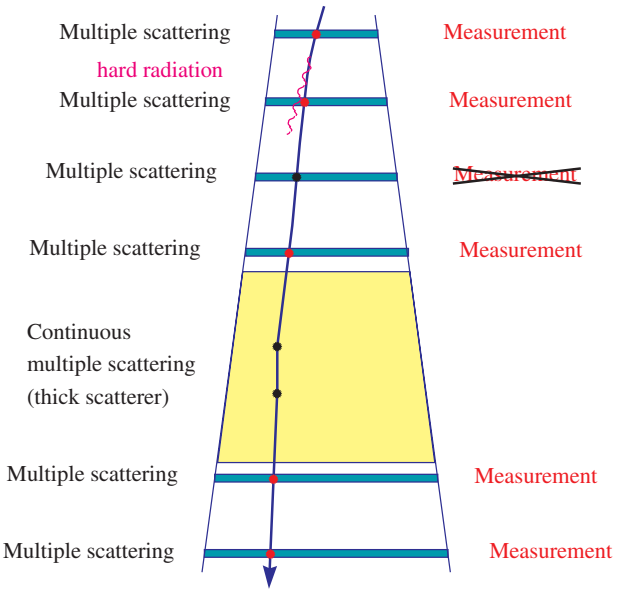

Figure 41. Schematic view of the information flow in the track fit. Elements shown are measurements (hits) in the tracking layers, in one case a missing hit which is, e.g. not found by the pattern recognition procedure, and random perturbations like multiple scattering and photon radiation.

$t_y = \tan\theta_y$ introduced in section 2.3.1. The stochastic nature of multiple scattering is that of a Markov process.

The distribution of the deflection angle follows a bell-like shape, though it cannot be accurately described by a Gaussian because of its pronounced tails. The variance of the projected multiple scattering angle is calculated within Molière theory [70–72] as

$$C_{\mathrm{MS}} = \left(\frac{13.6\,\mathrm{MeV}}{\beta pc}\right)^2 t[1 + 0.038\,\ln t]^2 \tag{45}$$

where t is the traversed path length in terms of radiation lengths x_R, usually called radiation thickness. (While the radiation length is frequently denoted by x_0 in the literature, the symbol x_R is used here instead to avoid confusion with other uses of x_0 throughout this article.) For a planar object arranged in a plane vertical to the z-axis, the radiation thickness along z is given by

$$\tilde{t} = \int \frac{\mathrm{d}z}{x_R(z)}. \tag{46}$$

Taking the track inclination against the z-axis into account, one obtains the effective radiation thickness

$$t = \tilde{t}\sqrt{1 + t_x^2 + t_y^2} \tag{47}$$

so that the final formula becomes (assuming $\beta \approx 1$)

$$C_{\mathrm{MS}} = \left(\frac{13.6\,\mathrm{MeV}}{pc}\right)^2 \sqrt{1 + t_x^2 + t_y^2}\,\tilde{t}\left[1 + 0.038\ln\sqrt{1 + t_x^2 + t_y^2}\tilde{t}\right]^2. \tag{48}$$

In general, multiple scattering could be treated in the track fit by expressing the angular uncertainty of each thin scatterer as an additional contribution to the error of each

affected measurement. Since a multiple scattering deflection will influence all downstream measurement errors in a correlated way, this introduces artificial correlations into the hitherto uncorrelated measurements, so that the matrix $\boldsymbol{V}$ in section 2.4.1 is no longer diagonal. Evaluation of equation (4) requires the inversion of non-trivial matrices whose dimension is not only the number of parameters but the number of measurements. Straightforward solutions of this problem have been devised [73], which intrinsically treat all multiple scattering angles as free parameters. In many practical situations, however, where the number of parameters may be five and the number of measurements perhaps as large as 70, this can lead to serious problems.

The generally accepted solution for the above problem is provided by the Kalman filter technique. The multiple scattering dilution is added as process noise (represented by the matrix Q_k in the transport equation, equation (8)) at the very position in the trajectory where it originates. The Kalman filter normally proceeds in the inverse flight direction along the path of the particle and takes the influences illustrated in figure 41 into account. Mathematically, the result will be identical to a straightforward least squares fit as described in the previous paragraph, but the detailed procedure avoids handling of huge matrices.

In Kalman filter language, the resulting covariance matrix contribution for thin scatterers is

$$\begin{aligned} \mathrm{cov}(t_x, t_x) &= (1 + t_x^2)(1 + t_x^2 + t_y^2)C_{\mathrm{MS}} \\ \mathrm{cov}(t_y, t_y) &= (1 + t_y^2)(1 + t_x^2 + t_y^2)C_{\mathrm{MS}} \\ \mathrm{cov}(t_x, t_y) &= t_x t_y \ \ (1 + t_x^2 + t_y^2)C_{\mathrm{MS}}. \end{aligned} \tag{49}$$

(These and related formulae and their derivation can be found in [74]).

It may be interesting to see how such a fit works in practice. In the following, we show the results of a study that has been performed on the basis of simulated events in the HERA-B geometry (figure 8), applying a Kalman filter based track fit to the simulated hits [75]. This kind of geometry is typical for modern forward spectrometers, and generally similar to COMPASS [76] or the planned LHCb [77] and bTEV [78]. The study was based on detector design resolutions and not intended to make quantitative statements on the actual spectrometer performance, but to provide insight into the effects of combining various different detector types, the sizeable number of hits per track, and the considerable amounts of material in the tracking area that make an accurate treatment of multiple scattering essential.

5.2.1. Impact parameter and angular resolutions. The visible track parameter resolution was obtained by calculating the track parameter residual for each track using the Monte Carlo truth, and applying a Gaussian fit to the distribution. (The term visible is used to distinguish this resolution from the one estimated by the fit.) The impact parameter resolution for tracks passing the Silicon micro-vertex detector and the outer tracker, as a function of momentum is shown in figure 42. Since this impact parameter is defined with respect to the position of the first hit of the track counting from the interaction point, the resolution is governed by the error of the first coordinate and only weakly dependent on momentum. Multiple scattering acts like a filter which dilutes the information from the following layers, and only at a higher momentum does their contribution to the resolution at the first point become visible.

Since the vertex detector measurement accuracy is approximately isotropic, horizontal and vertical resolution are almost identical; the deviation at $p = 100\,\mathrm{MeV}$ is explained by the fact that the strips in the first vertex detector layer are oriented almost parallel to the y-axis. The resolution of track slopes is shown in figure 43 and turns out to be dominated by the pronounced $\propto 1/p$ behaviour expected in a multiple scattering-dominated regime. At high momenta, the onset of coordinate resolution effects appears to be just visible, where the

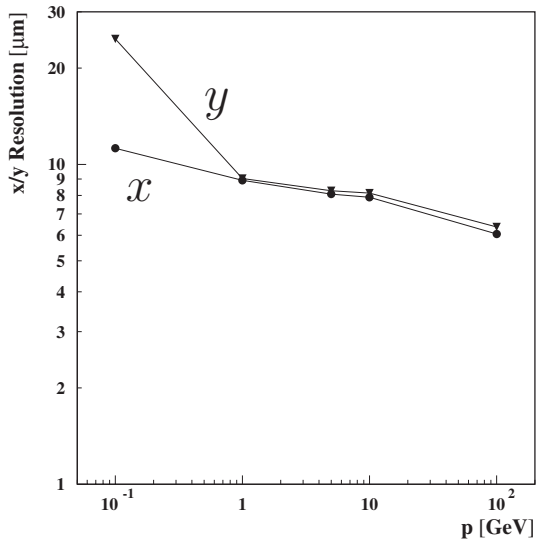

Figure 42. Impact parameter resolution at the first track point, separately for the coordinate in the bending plane (x, ●) and the non-bending plane (y, ▼).

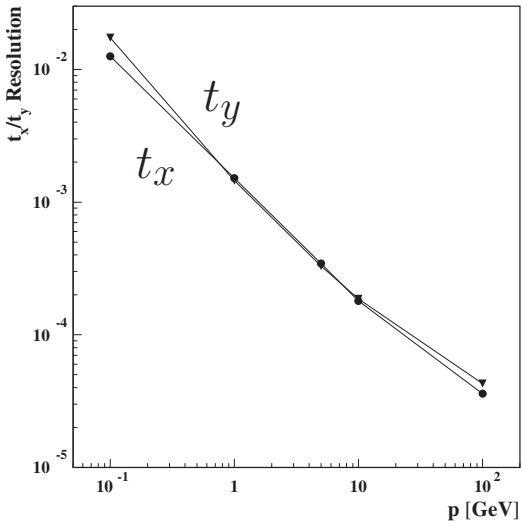

Figure 43. Resolution of the slope parameters $t_x = \tan\theta_x$ (●) and $t_y = \tan\theta_y$ (▼).

slightly better resolution of the horizontal slope (t_x) may be due to the predominantly vertical orientation (parallel to y) of the wires in the main tracking system.

The impact parameter resolution given above should not be confused with the quantities relevant for physics performance where assignment to vertices is important. In the latter case, the track parameters must be extrapolated from the first track point to the interaction area. With extrapolation distances of typically $\mathcal{O}(10\,\mathrm{cm})$, the resolution of the extrapolated impact parameters will generally be completely dominated by the angular resolution rather than the impact parameter resolution at the first point.

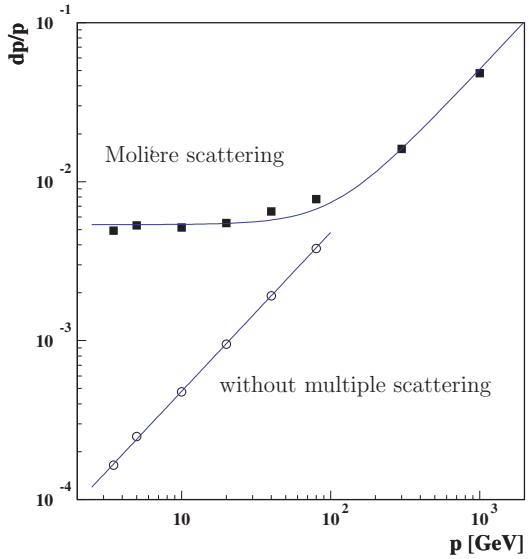

Figure 44. Visible momentum resolution (■) for simulated muons together with the fit of the parametrization described in the text (upper solid line). In the lower part, the open circles (○) show the relative momentum resolution with multiple scattering switched off, with a linear function fitted to it.

5.2.2. Momentum resolution. A very central design issue in spectrometers is the resolution of momentum, since it determines the rejection power against background in particle spectrometry. The relative momentum resolution, denoted by $\mathrm{d}p/p$, as a function of momentum is shown in figure 44 for particles traversing the areas SI, MC and PC of the spectrometer (see figure 8 for definition) in the polar angle area $0.1 < \theta < 0.15$. The circles show the relative momentum resolution that results with multiple scattering switched off in the simulation, leading to a strictly linear dependence on p. This behaviour is expected since the resolution is then only determined by the coordinate resolution and the geometrical layout of the spectrometer—size and number of layers—which provides the leverage for momentum measurement together with the magnetic field. The result reflects the fact that the curvature κ, which is the inverse of the radius of curvature, can be measured with a precision that is independent of its actual value; hence $\delta\kappa = \text{const}$. On the other hand, the curvature is inversely proportional to the momentum, so that $\mathrm{d}p/p \propto p$. In the presence of multiple scattering, the resolution shows a multiple scattering-dominated regime below momenta of $\approx$50 GeV, and a transition into a linear rise at high momentum. We have superimposed a fit with a constant and a linear resolution term added in quadrature. This parametrization, which corresponds to a commonly used function introduced by Gluckstern [7] for even spacing of tracking stations, does not fit the visible resolution very well in the momentum mid-range, which can be attributed to the uneven distribution of measurements, resolutions, material and magnetic field strength in the spectrometer.

5.2.3. Effects of fit non-linearity. The presence of the inhomogeneous magnetic field introduces particular effects of non-linearity into the fitting problem. The least squares fit technique, which the Kalman filter is built on, can still be applied, with the transport matrices now obtained as derivatives of the transport function. As already noted in section 2.4.1, the

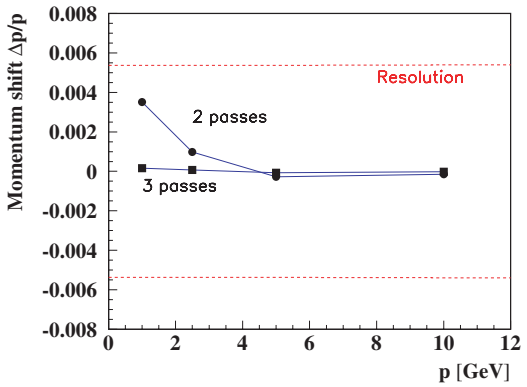

Figure 45. Residual of the momentum parameter (see equation (23)) normalized to the momentum itself for two and three passes of the fit. The relative momentum resolution is indicated by the dashed lines for comparison.

optimal properties of the least squares technique are still retained on the condition that the derivatives are taken at the position of the final trajectory. Since this is initially not necessarily the case, the fit must be repeated iteratively until the procedure converges.

The practical implications of non-linearity are visible in figure 45, which shows the mean relative deviation of the reconstructed momentum from the true value. Small systematic shifts of reconstructed momentum are observed for momenta below 5 GeV with two fit passes applied. These shifts reflect the convergence behaviour of the fit due to non-linearity. They are found to be virtually removed when a third pass is applied.

5.2.4. Contributions of different parts of the spectrometer. To understand detector design, it is also important to investigate how much different parts of the spectrometer contribute to the momentum measurement. In the HERA-B geometry (figure 8), the tracking system is grouped into the vertex detector (SI), the chambers within the magnet (MC), the chambers just behind the magnet (PC) and the so-called trigger chambers (TC), which are separated from the PC part by the RICH detector. In order to separate the contributions of the different spectrometer parts, the range of the fit was modified by omitting the vertex detector hits (labelled MC–PC range) and by adding the hits from the tracking chambers at the end of the main tracking system (SI–TC range). The resulting momentum resolutions are displayed in figure 46. It turns out that without including the vertex detector (MC–PC), the momentum resolution is well described by a constant and a linear term added in quadrature. In the regime of linear rise, the poorer coordinate resolution is reflected in comparison to the system including the vertex detector. When the fit, on the other hand, is extended into the 'TC' region, which is mainly designed to support the trigger (SI–TC), these additional measurements with their huge lever arm are expected to improve the coordinate contribution of the resolution. Such an improvement is visible in figure 46 for $p \geqslant 100$ GeV, where it is hardly relevant for the physics scope of the experiment. A third term proportional to the square-root of the momentum had to be added in quadrature to fit the resolution for the latter two ranges.

5.2.5. Parameter covariance matrix estimation. A very important task of the track fit is the quantification of the covariance matrix of the estimated track parameters. The reliability of parameter error estimation can be studied by investigating distributions of normalized

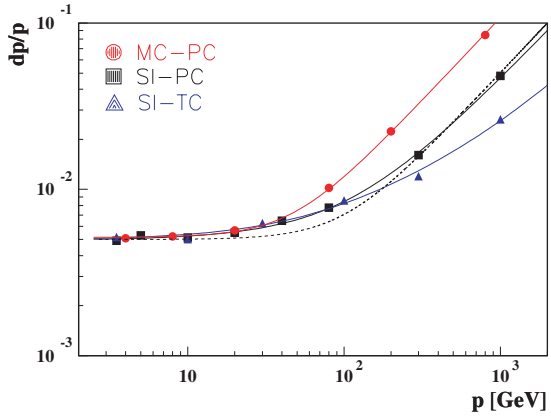

Figure 46. Relative momentum resolution in the MC–PC (circles), SI–PC (squares) and SI–TC (triangles) spectrometer ranges, together with the fits described in the text. The dashed line is the upper fit in figure 44.

parameter residuals (see equation (23) in section 2.5.5), which use the estimated error for normalization. In the example at hand, the resulting pull distributions are shown in figure 47, where unbiased fits with a Gaussian function are superimposed. Distortions of the parameter estimates would show up as deviations of the mean values from zero, which are, however, not present in this case. The Gaussian cores of the pulls agree in all cases with unity width, indicating a reliable estimate of the covariance matrix. One should note that only the mean value and variance of the pull distribution are indicators of the quality of the estimate. The actual shape of the distribution, e.g. whether it is Gaussian or not, reflects the underlying structure of the problem, as will be more clearly visible in the next section.

5.2.6. Goodness of fit. Since the Kalman filter is mathematically equivalent to a least-squares estimator, the sum of the filtered χ^2 contributions will follow a χ^2 distribution, provided that the random variables entering into the fit have Gaussian distributions. In this case, the χ^2 probability,

$$P_{\chi^2} = \int_{-\infty}^{\chi^2} f(\tilde{\chi}^2)\,\mathrm{d}\tilde{\chi}^2$$

where $f(\tilde{\chi}^2)$ is the standard χ^2 distribution for the appropriate number of degrees of freedom, should be evenly distributed between 0 and 1. (P_{χ^2} is often called the confidence level.) This prerequisite is not strictly fulfilled in the case of Molière scattering, so that deviations are to be expected. These effects have a potentially large influence in modern radiation hard drift chambers, where the drift cells are enclosed in a multitude of small gas volumes and a considerable amount of material is introduced into the tracking area.

Figure 48 compares the distribution of the χ^2 probability for the Gaussian form of multiple scattering (*a*) and Molière scattering (*b*). The peak at small probabilities in (*b*) obviously does not indicate a bad behaviour of the fit, but instead shows the inadequateness of the χ^2 test with non-Gaussian random variables. The probability distribution for various momentum values is shown in figure 49. The increasing prominence of the peak at low probability is clearly seen with decreasing momentum. Small χ^2 probability does not necessarily imply a bad estimation of the parameters, hence special care is required when a χ^2 cut is to be used to eliminate improperly reconstructed tracks.

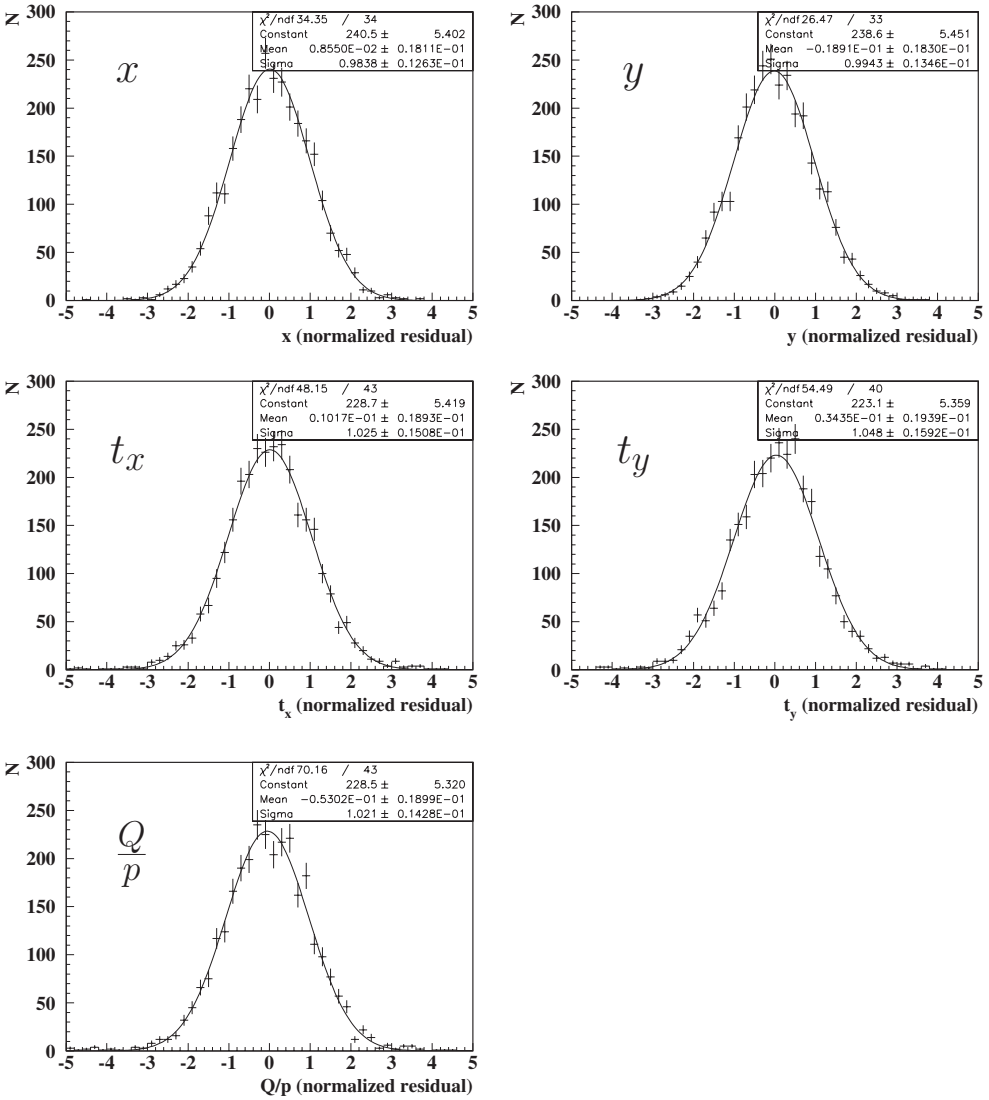

Figure 47. Normalized parameter residual distributions for muons of 10 GeV, based on 3000 simulated tracks.

5.3. Treatment of ionization energy loss and radiation

5.3.1. Ionization energy loss. For minimal ionizing particles in the gigaelectronvolts energy range, energy loss due to ionization within the tracking system depends in good approximation only on the amount of material that is traversed. In this case, it is not the radiation thickness (as defined in equation (46)), but the geometrical thickness multiplied by the mass density of the material that is relevant. Since the energy loss depends only weakly on the energy itself in this range, the effect will become most noticeable for low momentum particles. This behaviour is illustrated in figure 50, which shows the normalized residual of the momentum parameter Q/p for μ^+ particles of 3.5 and 10 GeV with ionization energy loss simulation turned on. The residual distributions are shifted towards positive values of Q/p, reflecting

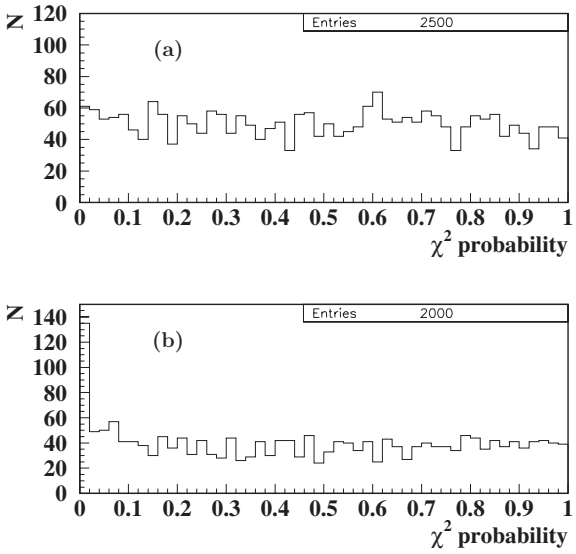

Figure 48. Distributions of χ^2 probability (confidence level) for the track fit for a 10 GeV particle: (*a*) with Gaussian form of multiple scattering; (*b*) with Molière scattering.

an underestimation of the energy, which is caused by the ionization energy loss, in particular upstream of the magnet. The visible shift corresponds to an energy loss of 12 MeV. On the other hand, the width of the residual distributions is not significantly increased, which in the 10 GeV case can be seen directly by comparing with figure 47.

A correction can be applied in each filter step if the value of $\mathrm{d}E/\mathrm{d}x$ of the particle in the material is known, since

$$E_{\text{after}} = E_{\text{before}} - \left(\frac{\mathrm{d}E}{\mathrm{d}x}\right)_{\text{ion}} \cdot \ell \tag{50}$$

where ℓ is the traversed thickness of the material. In general, this requires the knowledge of the particle mass. Since ionization energy loss will be most notable for small particle energies where the resolution is governed by multiple scattering, no correction to the momentum error has been applied. The bottom part of figure 50 displays the same normalized residuals with the energy loss correction applied. The bias of the momentum estimate is successfully eliminated by the correction.

5.3.2. Radiative energy loss. The corrections discussed up to now are usually sufficient for minimum ionizing particles. For electrons[5], however, the situation is more complicated since above the critical energy, which is of the order of millielectronvolts, these particles lose more energy through radiation of photons than through ionization when they traverse any material. This process is also of a more notably stochastic nature than ionization energy loss, as considerable fractions of the electron energy can be transferred to the photon. Modern radiation-hard detectors, such as, e.g., those under construction for the LHC, are confronted with this problem to a much higher degree than traditional detectors, because of the significant amount of material in the tracking system, which can easily exceed 50% of a radiation length.

[5] In this section the term electron should be interpreted to imply positron as well.

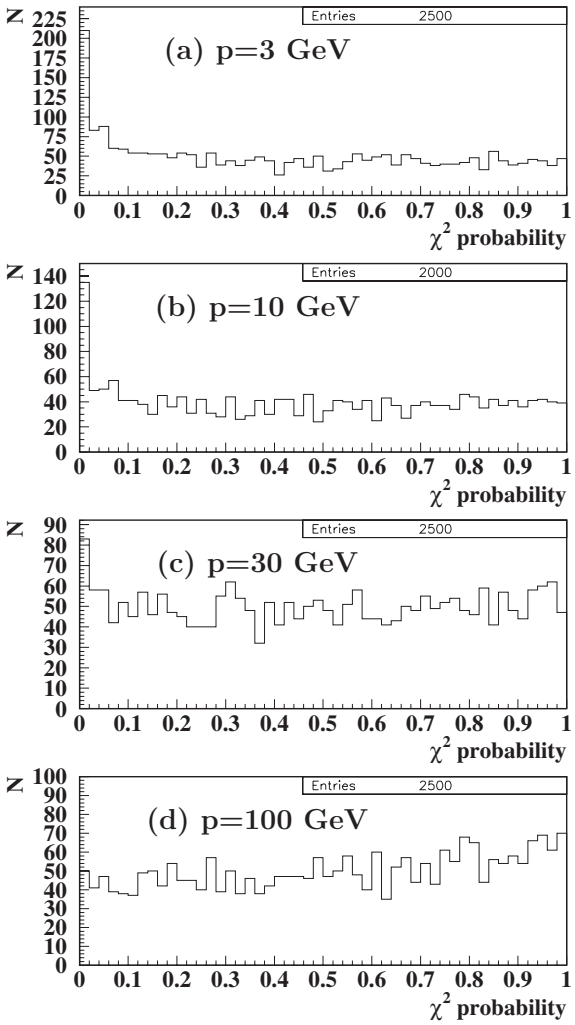

Figure 49. Distribution of χ^2 probabilities as a function of momentum: (*a*) 3.5 GeV, (*b*) 10 GeV, (*c*) 30 GeV and (*d*) 100 GeV.

For the relevance of photon radiation on measurement of the electron track parameters, three cases have to be distinguished regarding the range where the radiation occurs (indicated as regions 1–3 in figure 51).

Region 1: between the interaction point and the spectrometer magnet. If the point of origin of the particle is not yet within the magnetic field—as is typical for fixed-target setups rather than for collider detectors—radiation will not change the electron trajectory and thus not interfere with the quality of the fit; however, the spectrometer will only measure the remaining momentum of the electron after the radiation.

Region 2: within the magnetic field. In this case, the curvature of the trajectory changes because of the radiation, which means that the energy change is—in principle—measurable. Ignoring the radiation in the fit will lead to a bad description of the trajectory and to distortions of the parameter estimates.

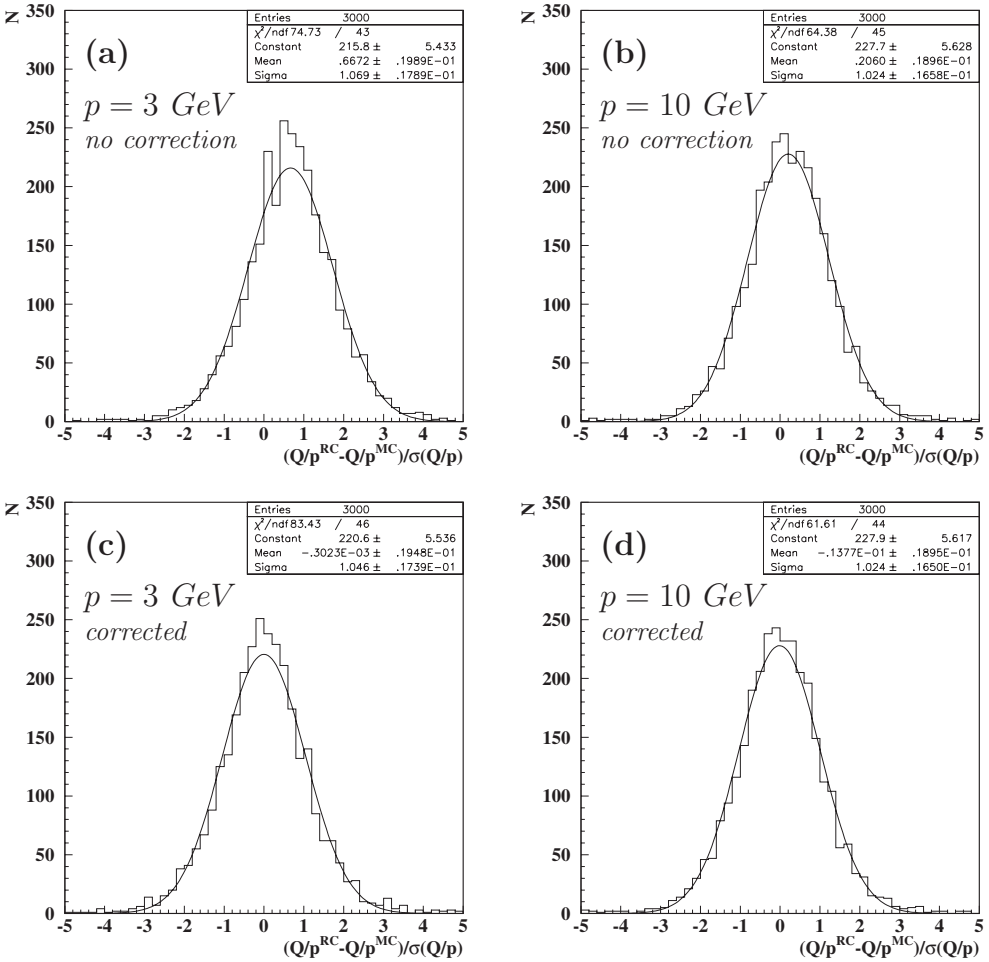

Figure 50. Pull distribution of the momentum parameter for μ^+ particles of 3 GeV (*a*) and (*c*) and 10 GeV (*b*) and (*d*). The upper pictures show the effect of $\mathrm{d}E/\mathrm{d}x$ if no correction is applied. The lower plots show the same when the correction is applied in the fit.

Region 3: beyond the magnetic field. If the electron loses energy downstream of the magnet, this will have no influence on the momentum measurement in the spectrometer. However, pair creation from radiated photons may lead to accompanying particles that can disturb pattern recognition in the downstream area.

The dilution due to energy loss of electrons and positrons through emission of electromagnetic radiation can be treated by the method proposed by Stampfer *et al* [79]. According to the Bethe–Heitler equation [80], this energy loss is described by

$$\left(\frac{\mathrm{d}E}{\mathrm{d}x}\right)_{\mathrm{rad}} = \frac{E}{x_R} \tag{51}$$

where x_R is the radiation length of the traversed material (see section 5.2). This leads to the relation

$$\left\langle \frac{E_{\mathrm{after}}}{E_{\mathrm{before}}} \right\rangle = \mathrm{e}^{-t} \tag{52}$$

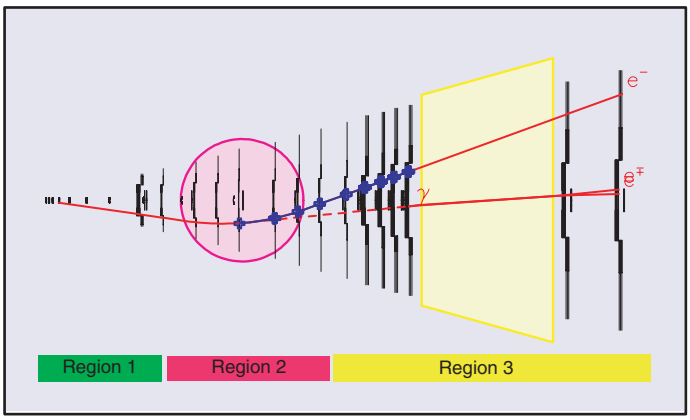

Figure 51. Regions 1–3 for classifying radiative energy loss illustrated in the geometry of the HERA-B spectrometer. The simulated geometry differs in some details from the one in figure 8. Also, the trajectory of a simulated electron is shown, which radiates a photon within the magnet, which is converted into a e^+e^- pair further downstream.

where t is the traversed distance measured in radiation lengths as defined before. For a track propagation which follows the track opposite to its physical movement, one obtains on average

$$\left(\frac{Q}{p}\right)' = \frac{Q}{p} + \Delta\left(\frac{Q}{p}\right) = \frac{Q}{p} - \frac{Q}{p}\frac{E_{\text{before}} - E_{\text{after}}}{E_{\text{before}}} = \frac{Q}{p}\mathrm{e}^{-t}. \tag{53}$$

The contribution to the propagated covariance matrix emerges as

$$\Delta\text{cov}\left(\frac{Q}{p}, \frac{Q}{p}\right) = \left(\frac{Q}{p}\right)^2 (\mathrm{e}^{-t(\ln 3/\ln 2)} - \mathrm{e}^{-2t}). \tag{54}$$

This contribution can be included into the Kalman filter process noise as introduced in equation (8).

5.3.3. Radiation energy loss correction within the magnetic field. Energy loss through radiation can not only interfere with the momentum measurement, but may also affect other track parameters. This is shown in figures 52(*a*), (*c*) and (*e*) which display the residuals of the parameters x, t_x and $1/p$ for electrons produced with 100 GeV momentum, where the fit was restricted to the magnet area (MC). Without the bremsstrahlung correction, the track slope estimate t_x shows a tail towards overestimated values, which is reflected in an underestimation of the corresponding impact parameter, x. The explanation for this effect is illustrated in figure 53 which for simplicity assumes a homogeneous field: the curvature of the electron track is abruptly increased beyond the point of radiation. Fitting the track with a constant momentum leads to an intermediate curvature resulting in a shift in the measured initial track slope.

The residual distribution of the momentum parameter, $1/p$, displays a tail towards higher values, corresponding to a mean momentum shift of $\approx$13%.

Also, the parameter errors are underestimated, which is evident from the normalized residuals in figures 54(*a*), (*c*) and (*e*) (uncorrected case), where the widths of the t_x and Q/p pull distributions are significantly enlarged.

Figures 52(*b*), (*d*) and (*f*) show the result with the radiation correction applied in the fit. One can see that the tails in the parameter estimates of x and t_x are far less pronounced, and the

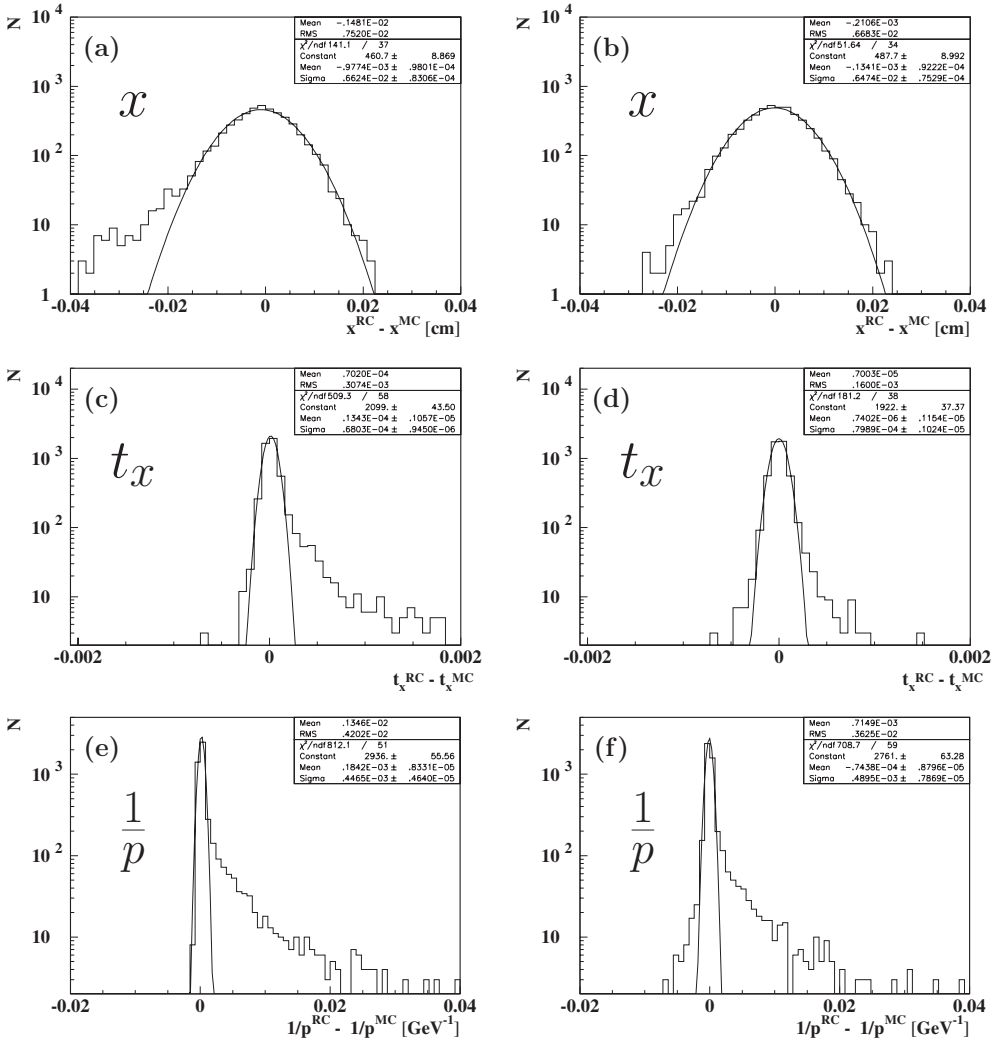

Figure 52. Distributions of parameter residuals for 100 GeV electrons based on 5000 tracks, where x is the impact parameter in the bending plane, $t_x = \tan\theta_x$ is the corresponding track slope, and $1/p$ the inverse momentum. The track fit was applied to all hits within the spectrometer magnet (region 2 in figure 51).

bias in the impact parameter and track slope is considerably reduced. Also, the distortion of the mean reconstructed inverse momentum $\delta(1/p) \approx \delta p/p^2$ is reduced from 1.3×10^{-3} to $7 \times 10^{-4}\,\mathrm{GeV}^{-1}$, and the standard deviation (RMS width) of the parameter estimates is reduced by 11% (x), 48% (t_x) and 14% (Q/p), respectively. Moreover, the radiation correction brings the RMS widths of the pull distributions close to unity (figures 54(b), (d) and (f)), which indicates a reliable covariance matrix estimate. The fit probability distribution is shown in figure 55. It reflects a non-χ^2 type distribution of the goodness-of-fit, which is expected since the bremsstrahlung radiation introduces a strongly non-Gaussian random perturbation.

The situation is different if one attempts to extend the radiation correction to the full tracking system including regions 1 and 3 which are outside of the magnetic field, most notably

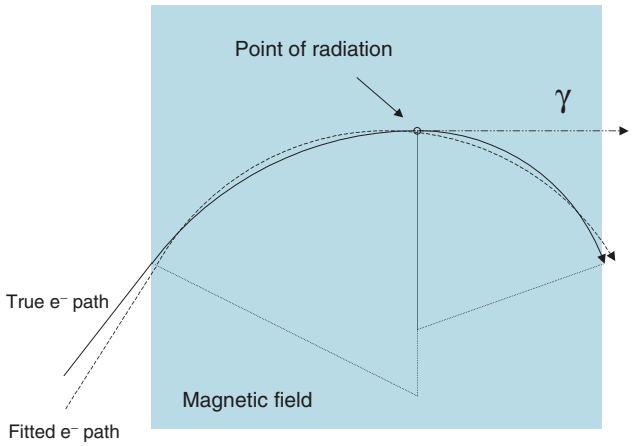

Figure 53. Illustration of how radiation within the magnetic field can affect the estimate of the fitted track slope. The magnetic field vector is pointing into the drawing plane. The electron, whose true path is shown by a solid line, emits a photon, which leads to an increase of curvature for the subsequent part of its trajectory within the magnet. This is illustrated by the curvature radii of the helices as dotted lines. The fitted trajectory (- - - -) assumes a single curvature, which leads to an overestimation of the initial slope of the track. The curvatures drawn are intentionally exaggerated.

the vertex detector whose material causes a significant energy loss for electrons. Outside of the magnetic field, however, the trajectory shape is not modified by radiation, which means that the fit will only apply the on-average correction according to the traversed radiation thickness. This can lead to bizarre results as seen in figure 56, which shows the distribution of the $1/p$ parameter residual multiplied by the momentum itself as well as the corresponding pull distribution. The peak has moved away from zero to negative residual values, implying that electrons in the peak obtain an overcorrected energy value. In figure 56(*b*), the mean value of the pull is close to zero, and the RMS width is close to one, indicating that the compensation works correctly in the statistical sense. For intuitive plausibility, however, it is relevant that a large fraction of measurements is in the immediate vicinity of the quoted value. A test of this criterion is shown in table 3, which summarizes the fraction of fits with momentum deviation of within 10% or 20% of the real value for the three correction scenarios. With the 10% criterion, the full spectrometer correction appears worse than even in the uncorrected case, while the correction restricted to the magnet gives the best description in the intuitive sense. In conclusion, the magnet-based correction appears to provide the best compromise, though this will, in general, have to be evaluated in each specific application.

5.4. Robust estimation

The preceding sections have shown how intrinsically non-Gaussian influences, such as multiple scattering, or radiative energy loss of electrons, can complicate the estimate of essential kinematic parameters and their interpretation. A fully adequate treatment of profoundly non-Gaussian variables is, in general, beyond the capabilities of least squares estimation. Likelihood methods, on the other hand, are in principle able to cope with random variables of any distribution, but often cannot be used with as efficient a machinery, in particular when it comes to computation of error matrices.

During the last years, promising concepts have been developed that permit the treatment of non-Gaussian random variables, but still allow us to use much of the powerful machinery

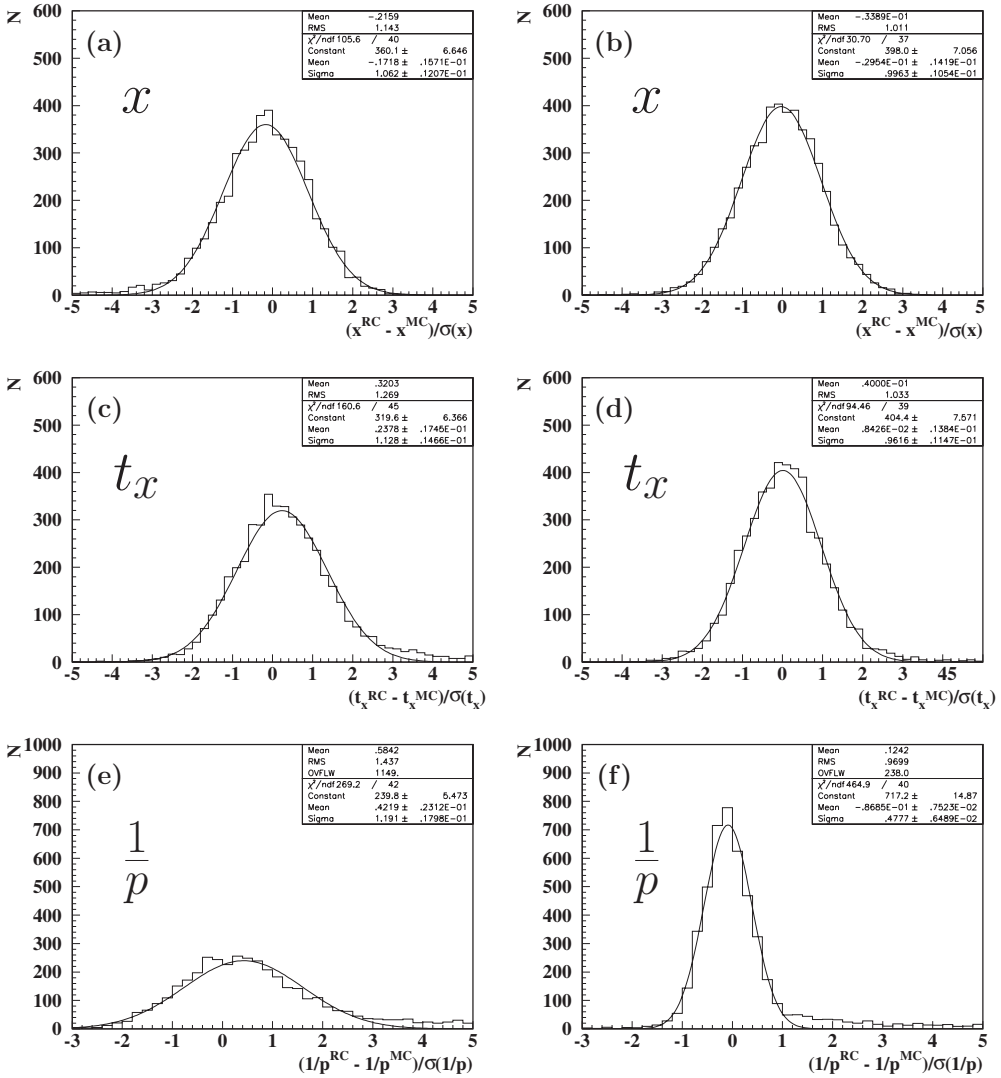

Figure 54. Distribution of normalized parameter residuals (pulls) for 100 GeV electrons based on 5000 tracks, where x is the impact parameter in the bending plane, $t_x = \tan\theta_x$ is the corresponding track slope, and $1/p$ the inverse momentum. The track fit was applied to all hits within the spectrometer magnet (region 2 in figure 51).

developed with least squares estimation. These methods are called robust estimation techniques. One very attractive idea is based on the fact that non-Gaussian distributions can often be approximated as a superposition of a limited number of Gaussian distributions [81,82]. For example, a distribution resembling a Gaussian in the centre, but featuring long tails, as is common with multiple scattering, can be approximated by a sum of a narrow Gaussian distribution and a wide one. If one performs two parallel least squares estimates, each based on one of the Gaussians, the resulting parameter estimates, combined with appropriate weights, will reflect the underlying statistics better than a single estimate with a single Gaussian approximation. Thus, the occurence of random variables in the tail of the distribution does not

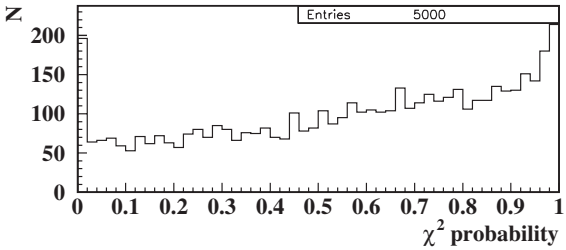

Figure 55. Distribution of χ^2 probability from the track fit of 100 GeV electrons in the main tracker with radiation correction. The non-Gaussian distribution of the radiated energy leads a non-flat probability distribution with a sharp peak near zero.

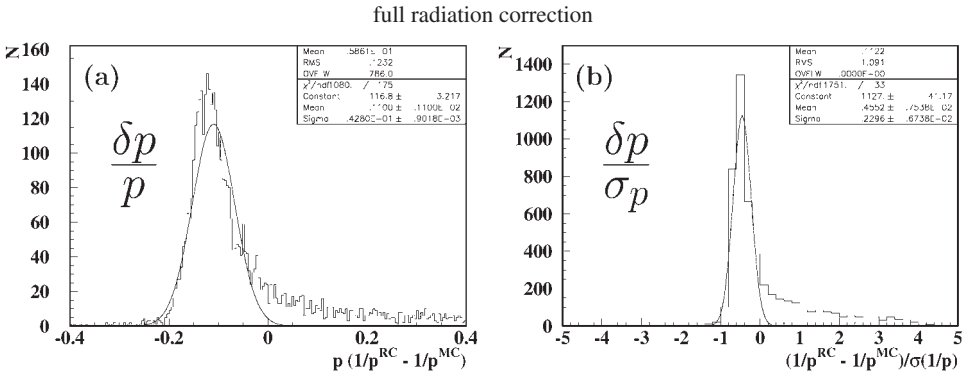

Figure 56. Distribution of $1/p$ parameter residuals multiplied by p as the measure for relative momentum deviation (a), and of normalized $1/p$ residuals (b) for 100 GeV electrons in the full tracker. The radiation correction was applied in the whole tracking system, leading to the shift of the peaks described in the text.

Table 3. Fraction of fits within given limits of momentum deviation, for three variants of radiation correction.

	Fraction of fits within momentum deviation	
Radiation correction mode	$-0.1 < \delta p/p < +0.1$	$-0.2 < \delta p/p < +0.2$
None	0.566 ± 0.004	0.678 ± 0.003
Within magnet	0.635 ± 0.003	0.728 ± 0.003
Within full spectrometer	0.321 ± 0.003	0.786 ± 0.002

pull the estimate as far away as it would with a traditional least squares estimator, leading to a more robust behaviour of the fit.

This is the basic idea of the Gaussian sum filter (GSF) [81–85], which uses the Kalman filter to incorporate the individual Gaussian components. Upon each occurrence of process noise, the distribution of which is approximated by a sum of N Gaussians, the filter splits into N parallel branches each of which obtains a corresponding weight. In a detector geometry with many scattering elements, this will lead to a repeated multiplication of the number of linear filters to be evaluated. To avoid enormous computing effort, the number of parallel components is limited by collapsing or clustering components of similar shape. It has been shown that the algorithm can be designed such that the computing effort increases linearly with the maximum number of parallel components (M), and that $M \approx 6$–8 already gives good

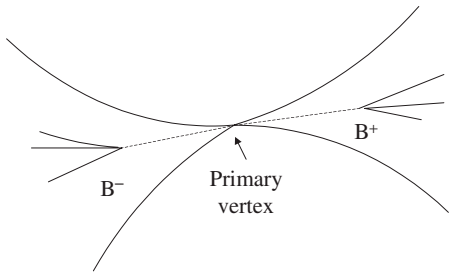

Figure 57. Schematic view of the event structure in an interaction of the type $e^+e^- \to B^+B^- + X$.

results [84]. In a similar way, radiative energy loss of electrons can be treated by approximating the radiated energy distribution by a superposition of several Gaussians [86].

6. Event reconstruction

After particle tracks have been reconstructed, they form the basis for the reconstruction of the whole event. This will ultimately include particle identification based on $\mathrm{d}E/\mathrm{d}x$, time-of-flight, Čerenkov or transition radiation, muon chambers and calorimetry, as well as kinematical reconstruction of composite particles and jets. This paper will restrict itself to a brief discussion of vertex reconstruction and kinematical constraints.

6.1. Vertex pattern recognition

The vertex is an essential element of the space-time structure of an interaction. Vertices indicate either the location where an interaction has taken place, for example, the primary interaction that is the ultimate origin of all emerging particles, or the place where an unstable particle has decayed. This is illustrated in figure 57, which schematically sketches the final state of an interaction with the associated production of two beauty mesons, as can occur for example, at a high energy e^+e^- collider. The beauty hadrons, here a B^+ and a B^-, are produced together with accompanying charged particles at the interaction point, travel invisibly for some distance, which is, on average, determined by their lifetime and momentum, whereupon they decay into daughter particles. The charged tracks coming from these decays can be used to reconstruct the decay locations of the B mesons as secondary vertices[6]. The other tracks, together with the reconstructed B mesons form the primary vertex, which indicates the interaction point.

In many practical applications, the vertex is constructed by an iterative procedure as illustrated in figure 58. In most cases, some *a priori* knowledge about the vertex position exists, for example, the shape of the beam spot, in which interactions occur in the first place. Then, a first track is selected as a vertex seed, which already narrows down the covariance ellipsoid in two dimensions. When a second suitable track is added, the vertex is already closely defined in all coordinates. This provides strong rejection power against off-vertex particles when adding more tracks.

As in the track pattern recognition case, the danger lies in the dependence on the starting point. It is, therefore, necessary to use iterative criteria which ensure that the track forming the

[6] We neglect here the complication that the B meson is likely to decay to a final state with a charmed particle, which again has a non-negligible lifetime.

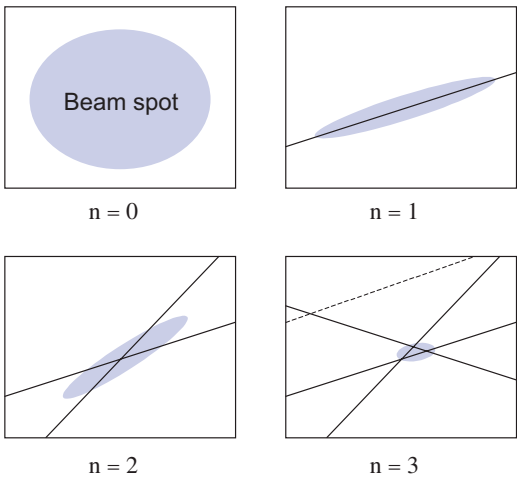

Figure 58. Illustration of the iterative construction of a (primary) vertex, where n is the number of tracks used to define the vertex in each step. The shaded area indicates the covariance ellipse of the projected vertex after each step. The dashed line indicates an outlier track.

vertex seed is well chosen, and even then it must be possible to scrutinize the track ensemble of a vertex, to remove tracks that have turned out to be off the mark, and to reconnect tracks that have been discarded at an earlier stage of the construction. The vertex algorithm used in the ZEUS experiment [87], which internally uses the fitting methods of [88], may serve as an example: it uses the proton beam line as a soft constraint, and then produces a set of all track pairs that would be compatible with a common vertex together with the beam line constraint within a suitable χ^2 margin. The track pairs are then ordered according to their degree of compatibility with other track pairs, defined by the criterion above. The track pair of highest compatibility then forms the first vertex seed to be used, though other track pairs of high compatibility level are also tried, and in the end the best set is chosen based on a criterion based on the number of tracks and total χ^2. Other approaches start by connecting all tracks to a diffuse master vertex, which is then successively split into vertices of smaller multiplicities and isolated tracks. A systematic investigation of different methods for vertex reconstruction in the context of the CMS experiment can be found in [89].

An entirely different approach is pursued in the topological vertex finding algorithm [90] developed for the vertex detector of the SLD experiment [17]. This method assigns a Gaussian tube around each track extrapolation to indicate the likelihood of an assigned vertex on a single track basis. The Gaussian tubes of all tracks are then combined to find points with maximum probability of a vertex. This method resembles the fuzzy radon transform for tracks discussed in section 3.2. The search for maxima is then performed by sophisticated clustering algorithms. A particularly intriguing feature is the efficient resolution of heavy flavour cascade decays.

Direct vertex search by the Hough transform is possible in cases where the vertex location is already strongly constrained in some coordinates, for example, through the shape of a wire target [91].

6.2. Vertex fitting

The least-squares principle can also be readily applied for vertex fitting [92–94]. The parameters of the tracks $\vec{p}_1, \ldots, \vec{p}_n$ at a given reference surface plus the *a priori* knowledge

of the vertex are the input, and the calculated vertex position together with the reduced track parameters of each particle, which contain only directional and momentum information at the common vertex, are the output. A general property of vertex fitting is the fact that, unlike track fitting, the fit is always non-linear, since even with straight-line tracks the extrapolation to the vertex introduces a coupling between positional and directional parameters.

As noted earlier, already vertex pattern recognition requires incremental, progressive fitting, with tracks added or removed one by one. It is, therefore, not surprising that, the Kalman filter is in many cases the method of choice [95] for vertex fitting also. In the vertex fitting case, the transport becomes trivial, and also process noise does not have an equivalent. The filter step adds another track to the vertex and updates the vertex position as well as the reduced track parameters. It is very easy to remove an already filtered track from the vertex candidate, since in the filter equations, the inverse covariance matrix of the track acts as the weight of the track information, and setting its sign to negative will subtract the track from the vertex fit. We prefer not to display the Kalman filter equations for vertex fitting here explicitly, but refer to the literature [27].

6.3. Kinematical constraints

Pattern recognition deals with merging of measured information with *a priori* knowledge. For example, in track pattern recognition the track model enhances the measurement power of each individual hit, while vertex assignment improves the spatial information of each associated track. In similar fashion, *a priori* knowledge can be used in many cases in the further reconstruction of the event. A typical example is the beam energy constraint: in e^+e^- b-physics experiments which operate at the $\Upsilon(4S)$ energy, as BaBar, BELLE, CLEO and the earlier ARGUS, the B mesons are produced in an exclusive decay of the $\Upsilon(4S)$ resonance, and the energy of the B mesons is precisely the beam energy, which is known to a much better precision than the B meson energy reconstructed from its measured decay particles. Imposing the beam energy constraint, then, also improves the resolution of the B candidate mass; this method has been a vital tool in the investigation of exclusive B decays (see, e.g. [96]).

Also, masses of intermediate particles in a decay chain, for example, $B^0 \to D^{*+}\pi^+\pi^-\pi^-$, $D^{*+} \to D^0\pi^+$, $D^0 \to K^-\pi^+$ can be used to imply kinematical constraints. In this case, the D^0 is a rather stable particle whose width is too small to resolve by direct kinematical reconstruction in a spectrometer. Therefore, the established knowledge of the D^0 mass [97] can be imposed as a kinematical constraint. For example, if $\vec{\alpha}$ denotes the reconstructed parameters of the K^- and π^+ particles and V_α their covariance matrix, the reconstructed D^0 mass will be a function $M(\alpha)$ of these parameters, and the introduction of a Lagrange multiplier μ leads to the expression

$$X^2 = (\vec{\alpha}_c - \vec{\alpha})^T V_\alpha^{-1} (\vec{\alpha}_c - \vec{\alpha}) + 2\mu(M(\vec{\alpha}_c) - m_{D^0}) \tag{55}$$

which has to be minimized with respect to the constrained parameters $\vec{\alpha}_c$. If the daughter particles form a secondary vertex, their parameters can be optimized as well. The D^0 mass constraint leads, in general, to a considerable improvement of the D^* mass peak, which becomes much narrower than the experimental resolution. In comparison to the popular mass difference method, which benefits from the correlation in the errors of the reconstructed D and D^* masses, this approach has the advantage that the result can be used in turn to reconstruct more complex decay chains of angular excitations in the D systems, or of B hadrons. In a next step of B reconstruction, even the tabulated D^* mass could be imposed as another independent constraint.

7. Concluding remarks

The variety of pattern recognition tasks in particle physics tracking detectors has led to a multitude of different approaches. Several of the global methods, such as template matching or Hough transform/histogramming, play an unchallenged rôle in special applications, while Hopfield networks and deformable templates frequently appear to be either limited to favourable scenarios (e.g. with 3D measurements and moderate occupancy), or need an excellent initialization or combination with a track following algorithm to become applicable at production scale. In the case of elastic arms, the choice of an efficient minimization technique is also essential. Local methods of pattern recognition are still going strong, with the Kalman filter as the mathematical backbone, and accompanied by subtle arbitration techniques they can cope well even with high track densities and sizeable amounts of material in the tracking area. The new generation of high energy hadron colliders, in particular the LHC with huge track densities in piled-up events will become an important benchmark for algorithm performance. It can be expected that sophisticated combination of both global and local approaches in different passes of the procedure, matched to the particular layout of each experiment, will become a promising path to achieving the best performance.

The increasing abundance of material in radiation hard detectors also poses additional challenges to track fitting. While the correction of multiple scattering with the Kalman filter has become the accepted general standard, Molière scattering tails require a careful interpretation of the results. Electron energy reconstruction with sizeable radiative energy loss is a major challenge and requires very careful treatment, and becomes a rewarding subject for robust methods beyond least squares estimation. Also, vertex pattern recognition can be expected to receive increasing attention in very complex event topologies at LHC, where reliable tagging of heavy flavour is a crucial prerequisite to scientific discovery.

Acknowledgment

It is a pleasure to thank E Lohrmann for his valuable comments on the manuscript.

References

[1] Grote H 1987 Review of pattern recognition in high energy physics *Rep. Prog. Phys.* **50** 473–500
[2] Albrecht H *et al* 1995 Search for rare B decays *Phys. Lett.* B **353** 554–62
[3] ATLAS Collaboration 1997 *ATLAS Inner Detector Technical Design Report* vol I, CERN/LHCC/97-16, CERN
[4] Grupen C 1996 Particle detectors *Cambridge Monographs on Particle Physics*
[5] Kleinknecht K 1999 *Detectors for Particle Radiation* (Cambridge: Cambridge University Press)
[6] Green D 2000 *The Physics of Particle Detectors* (Cambridge: Cambridge University Press)
[7] Gluckstern R L 1963 Uncertainties in track momentum and direction, due to multiple scattering and measurement errors *Nucl. Instrum. Methods* **24** 381–9
[8] Carlin R *et al* (ZEUS Collaboration) 2003 The ZEUS microvertex detector *Nucl. Instrum. Methods* A **511** 23–37
[9] Danilov M *et al* 1983 The ARGUS drift chamber *Nucl. Instrum. Methods* **217** 153–9
[10] Sciolla G *et al* (BaBar Collaboration) 1998 The babar drift chamber *Nucl. Instrum. Methods* A **419** 310–14
[11] Bruyant F, Lesceux J M and Plano R J 1980 The butterfly drift chamber geometry: an optimal four-plane drift chamber for use in a high track multiplicity environment *Nucl. Instrum. Methods* **176** 409
[12] Kind O *et al* 2003 A ROOT-based client-server event display for the zeus experiment *Proc. Computing in High Energy Physics Conf. (La Jolla, 2003) Preprint* hep-ex/0305095
[13] Križan P *et al* 1994 HERA-B, an experiment to study CP violation at the HERA proton ring using an internal target *Nucl. Instrum. Methods* A **351** 111–31
[14] Hartouni E *et al* 1995 HERA-B: an experiment to study CP violation in the B system using an internal target at the HERA proton ring *Design Report* DESY-PRC-95-01

[15] Mankel R 1998 The HERA-B experiment: overview and concepts *Proc. International Conf. on High-Energy Physics (ICHEP 98) (Canada, Vancouver, 1998)* vol 2, pp 1513–18
[16] Wieman H *et al* 1997 STAR TPC at RHIC *IEEE Trans. Nucl. Sci.* NS-44 671–8
Thomas J H 2002 A TPC for measuring high multiplicity events at RHIC *Nucl. Instrum. Methods* A **478** 166–9
Anderson M *et al* 2003 The STAR time projection chamber: a unique tool for studying high multiplicity events at RHIC *Nucl. Instrum. Methods* A **499** 659–78
[17] Taylor F E *et al* 1996 Design and performance of a 307 million pixel CCD vertex detector *Proc. 28th International Conf. on High Energy Physics (ICHEP 96) (Warsaw, 1996)* vol 2, pp 1739–42
[18] Brandt S 1992 *Datenanalyse* Bibliographisches Institut Mannheim (in German)
[19] Blobel V and Lohrmann E 1998 *Statistische und Numerische Methoden der Datenanalyse* (Leipzig: Teubner) (in German)
[20] Bevington P and Robertson D 1992 *Data Reduction and Error Analysis for the Physical Sciences* (New York: McGraw Hill)
[21] Eadie W T *et al* 1971 *Statistical Methods in Experimental Physics* (Amsterdam: North-Holland)
[22] Frodesen A G, Skjeggestad O and Tofte H 1979 *Probability and Statistics in Particle Physics* Universitetsforlaget
[23] Kalman R E 1960 A new approach to linear filtering and prediction problems *Trans. ASME J. Basic Eng.* **82** 35–45
Kalman R E and Bucy R S 1961 New results in linear filtering prediction theory *Trans. ASME J. Basic Eng.* **83** 95–108
[24] Billoir P 1984 Track fitting with multiple scattering: a new method *Nucl. Instrum. Methods* **225** 352–66
[25] Frühwirth R 1987 Application of Kalman filtering *Nucl. Instrum. Methods* A **262** 444–50
[26] Billoir P and Qian S 1990 Simultaneous pattern recognition and track fitting by the Kalman filtering method *Nucl. Instrum. Methods* A **294** 219–28
[27] Böck R H, Grote H, Notz D and Regler M 2000 *Data Analysis Techniques for High-Energy Physics Experiments* 2nd edn 1990 (with R. Frühwirth) (Cambridge: Cambridge University Press)
[28] Brown D N, Charles E A and Roberts D A The BaBar track fitting algorithm *Proc. Computing in High Energy Physics Conf. (Padova, 2000)*
[29] Mankel R and Spiridonov A 1999 *Compatibility Analysis* HERA-B Internal Note 99-111
[30] Schulz H D and Stuckenberg H J 1981 *Proc. Topical Conf. on Application of Microprocessors in High Energy Physics Experiments* CERN 81-07
[31] Koch N *et al* 1996 The ARGUS vertex trigger *Nucl. Instrum. Methods* A **373** 387–405
[32] Seidel S *et al* (ARGUS Collaboration) 1991 The ARGUS micro-vertex drift chamber *Proc. APS Conf. Particles and Fields (Vancouver, 1991)* vol 2, pp 1158–63
[33] Dell'Orso M and Ristori L 1990 A highly parallel algorithm for track finding *Nucl. Instrum. Methods* A **287** 436–8
[34] Battaiotto P *et al* 1990 The tree-search processor for real-time track pattern recognition *Nucl. Instrum. Methods* A **287** 431–5
[35] Ackerstaff K *et al* 1998 The HERMES spectrometer *Nucl. Instrum. Methods* A **417** 230–65
[36] Blom J *et al* 1994 A fuzzy radon transform for track recognition *Proc. Computing in High Energy Physics Conf. (San Francisco, 1994)*
[37] Gyulassy M and Harlander M 1991 Elastic tracking and neural network algorithms for complex pattern recognition *Comput. Phys. Commun.* **66** 31–46
[38] Antonov A 2003 (Moscow Engineering and Physics Institute) private communication
[39] Hough P V C 1959 Machine analysis of bubble chamber pictures *Int. Conf. on High Energy Accelerators and Instrumentation* CERN, pp 554–6
[40] Ohlsson M, Peterson C and Yuille A L 1992 Track finding with deformable templates: the elastic arms approach *Comput. Phys. Commun.* **71** 77–98
[41] Borgmeier C 1996 Global pattern recognition in the HERA-B tracking system *Diploma Thesis* Humboldt University Berlin (in German)
[42] Schober T 1996 Investigation of Hough transforms as global approaches to pattern recognition in the HERA-B main tracking system *Diploma Thesis* Humboldt University Berlin (in German)
[43] Anderson J A and Rosenfeld E 1988 *Neurocomputing: Foundations of Research* (Cambridge: MIT Press)
[44] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities *Proc. National Academy of Science, USA* vol 79, pp 2554–8 Reprinted in [43]
[45] Shrivastava Y, Dasgupta S and Reddy S M 1992 Guaranteed convergence in a class of Hopfield networks *IEEE Trans. Neural Netw.* **3** 951–61 Reprinted in [43]
[46] Denby B 1988 Neural networks and cellular automata in experimental high energy physics *Comput. Phys. Commun.* **49** 429–48

[47] Peterson C 1989 Track finding with neural networks *Nucl. Instrum. Methods* A **279** 537–49
[48] Stimpfl-Abele G and Garrido L 1991 Fast track finding with neural networks *Comput. Phys. Commun.* **64** 46–56
[49] Mankel R 1997 Pattern recognition algorithms for B meson reconstruction in hadronic collisions *Proc. Computing in High Energy Physics Conf. (Berlin, 1997)* URL http://www.ifh.de/CHEP97/paper/183.ps
[50] Abt I *et al* 2002 Cellular automaton and Kalman filter based track search in the HERA-B pattern tracker *Nucl. Instrum. Methods* A **490** 546–58
[51] Abt I *et al* 2002 CATS: a cellular automaton for tracking in silicon for the HERA-B vertex detector *Nucl. Instrum. Methods* A **489** 389–405
[52] Ohlsson M 1993 Extensions and explorations of the elastic arms algorithm *Comput. Phys. Commun.* **77** 19–32
[53] Peterson C and Söderberg B 1989 A new method for mapping optimization problems onto neural networks *Int. J. Neural Syst.* **1** 3–22
[54] Lindström M 1995 Track reconstruction in the ATLAS detector using elastic arms *Nucl. Instrum. Methods* A **357** 129–49
[55] Blancenbecler R 1994 Deformable templates—revised and extended—with an OOP implementation *Comput. Phys. Commun.* **81** 318–34
[56] Paus A 1997 Pattern recognition in the tracking system of the HERA-B detector with an elastic arms algorithm *Diploma Thesis* Humboldt University, Berlin (in German)
[57] Fahlmann S E 1988 Faster-learning variations on back-propagation: an empirical study *Proc. Connectionist Models Summer School (San Mateo, 1988)*
[58] Riedmiller A and Braun H 1993 A direct adaptive method for faster backpropagation learning: the RPROP algorithm *Proc. IEEE International Conf. on Neural Networks (San Francisco, 1993)*
[59] Frühwirth R and Strandlie A 1999 Track fitting with ambiguities and noise: a study of elastic tracking and nonlinear filters *Comput. Phys. Commun.* **120** 197–214
[60] Strandlie A and Frühwirth R 2000 Adaptive multitrack fitting *Comput. Phys. Commun.* **133** 34–42
[61] Mankel R 1997 A concurrent track evolution algorithm for pattern recognition in the HERA-B main tracking system *Nucl. Instrum. Methods* A **395** 169–84
[62] Albrecht H private communication
[63] Khanov A *et al* 2002 Tracking in CMS: software framework and tracker performance *Nucl. Instrum. Methods* A **478** 460–4
[64] Angarano M M *et al* (CMS Tracker Collaboration) 2003 The silicon strip tracker for CMS *Nucl. Instrum. Methods* A **501** 93–9
[65] Mankel R and Spiridonov A 1999 The concurrent track evolution algorithm: extension for track finding in the inhomogeneous magnetic field of the HERA-B spectrometer *Nucl. Instrum. Methods* A **426** 268–82
[66] Regler M, Frühwirth R and Mitaroff W 1996 Filter methods in track and vertex reconstruction *Int. J. Mod. Phys.* C **7** 521–42
[67] Mankel R 1996 Online track reconstruction for HERA-B *Nucl. Instrum. Methods* A **384** 201–6
[68] Press W H *et al* 1993 *Numerical Recipes in C: The Art of Scientific Computing* 2nd edn (Cambridge: Cambridge University Press)
[69] Oest T 1997 *Particle Tracing Through The HERA-B Magnetic Field* HERA-B Internal Note 97-165
[70] Bethe H A 1953 Molière's theory of multiple scattering *Phys. Rev.* **89** 1256–66
[71] Highland V L 1975 *Some practical remarks on multiple scattering Nucl. Instrum. Methods* **129** 497–9
Highland V L 1979 Erratum *Nucl. Instrum. Methods* **161** 171
[72] Lynch G R and Dahl O L 1991 Approximations for multiple coulomb scattering *Nucl. Instrum. Methods* B **58** 6–10
[73] Lutz G 1988 Optimum track fitting in the presence of multiple scattering *Nucl. Instrum. Methods* A **273** 349–74
[74] Wolin E J and Ho L L 1993 Covariance matrices for track fitting with the Kalman filter *Nucl. Instrum. Methods* A **329** 493–500
[75] Mankel R *The Object-Oriented Track Fit* HERA-B Internal Note 98-079
[76] Baum G *et al* (COMPASS Collaboration) *Common Muon and Proton Apparatus for Structure and Spectroscopy (Proposal)* CERN/SPSLC 96-14
[77] Amato S *et al* (LHCb Collaboration) 1998 *LHCB Technical Proposal* CERN-LHCC-98-4, CERN-LHCC-P-4
[78] Papavassiliou V *et al* (BTeV Collaboration) 2000 BTeV: a proposal for a new B physics experiment at the Fermilab tevatron collider (La Thuile 2000) *Results And Perspectives in Particle Physics* pp 843–64
[79] Stampfer D, Regler M and Frühwirth R 1994 Track fitting with energy loss *Comput. Phys. Commun.* **79** 157–64
[80] Bethe H A and Heitler W 1934 *Proc. R. Soc.* A **146** 83
[81] Kitagawa G 1989 Non-Gaussian seasonal adjustment *Comput. Math. Appl.* **18** 503–14
[82] Kitagawa G 1994 The two-filter formula for smoothing and an implementation of the Gaussian-Sum smoother *Ann. Inst. Statist. Math.* **46** 605–23

[83] Frühwirth R 1995 Track fitting with long-tailed noise: a Bayesian approach *Comput. Phys. Commun.* **85** 189–99
[84] Frühwirth R 1997 Track fitting with non-Gaussian noise *Comput. Phys. Commun.* **100** 1–16
[85] Frühwirth R and Regler M 2001 On the quantitative modelling of tails and core of multiple scattering by Gaussian mixtures *Nucl. Instrum. Methods* A **456** 369
[86] Frühwirth R and Frühwirth-Schnatter S 1998 On the treatment of energy loss in track fitting *Comput. Phys. Commun.* **110** 80–6
[87] Hartner G F 1998 *VCTRAK Briefing: Program and Math* ZEUS Internal Note 98-058
[88] Billoir P and Qian S 1992 Fast vertex fitting with local parametrization of tracks *Nucl. Instrum. Methods* A **311** 139–50
[89] Frühwirth R *et al* 2003 New vertex reconstruction algorithms for CMS *Proc. Computing in High Energy Physics Conf. (La Jolla, 2003) Preprint* physics/0306012
[90] Jackson D J 1997 A topological vertex reconstruction algorithm for hadronic jets *Nucl. Instrum. Methods* A **388** 247–53
[91] Lohse T 1995 *Vertex Reconstruction and Fitting* HERA-B Internal Note 95-013
[92] Saxon D H 1985 Three-dimensional track and vertex fitting in chambers with stereo wires *Nucl. Instrum. Methods* A **234** 258–66
[93] Forden G E and Saxon D H 1986 Improving vertex position determination using a kinematic fit *Nucl. Instrum. Methods* A **248** 439–50
[94] Saxon D H 1987 Vertex detection and tracking at future accelerators *Hadronic J.* **10** 117–39
[95] Luchsinger R and Grab C 1993 Vertex reconstruction by means of the Kalman filter *Comput. Phys. Commun.* **76** 263–80
[96] Albrecht H *et al* 1990 Exclusive hadronic decays of B mesons *Z. Phys.* C **48** 543–51
[97] Hagiwara K *et al* (Particle Data Group) 2002 Review of particle physics *Phys. Rev.* D **66**