Presented at the 5th International Linear Collider Workshop (LCWS 2000), Oct 24-28, 2000, Batavia, IL SLAC-PUB-10798

Clustering Algorithm Studies

Norman A. Graf

Stanford Linear Collider Center

Abstract. An object-oriented framework for undertaking clustering algorithm studies has been developed. We present here the definitions for the abstract Cells and Clusters as well as the interface for the algorithm. We intend to use this framework to investigate the interplay between various clustering algorithms and the resulting jet reconstruction efficiency and energy resolutions to assist in the design of the calorimeter detector.

ENERGY FLOW CONCEPT

The basic concept of the "Energy Flow" algorithm for jet finding is to use the tracking detectors for the measurement of charged particle momenta and the calorimeter for neutral particle energy measurement. It is therefore essential that one be able to identify energy deposition in the calorimeter arising from individual particles. The first prerequisite for this is to build a calorimeter with fine enough segmentation to separate the particle showers. The next is to be able to associate the hit cells in the calorimeter into clusters which can then be identified with the particles. It is the goal of this study to enable detector design to be intelligently driven by the reconstruction capabilities of a calorimeter in the Linear Collider physics and background environment.

Patterns of energy deposition in the calorimeter can be characterized by the type of particle initiating the shower and fall into three broad categories: electromagnetic showers, hadronic showers and minimum ionizing. The first class corresponds to electron and photon showers, which are very localized and highly correlated. The shower shape can be modeled very well with analytic formulae or by comparison to test beam or Monte Carlo simulations. The energy deposited in the calorimeter is also very well correlated to the energy of the particle itself. Muons interact minimally with the material in the calorimeter, depositing only minimum ionization, but do so predictably and along their entire trajectory. One can reconstruct the muon's path through the calorimeter just as one does through a tracking detector. Since the muons typically deposit a minimal amount of energy in the detector, systematic uncertainties due to a mismeasurement of the energy are small. Hadronic showers are the most difficult to generalize, since they are broad and tend to deposit energy in a disconnected fashion. Much of the energy is not directly reconstructible, leading to much poorer energy resolution. It is this observation which has lead to the "Energy Flow" algorithm of jet reconstruction.

Work supported by the Department of Energy contract DE-AC02-76SF00515.

Stanford Linear Accelerator Center, Stanford University, Stanford, California, 94309

The task, then, is to efficiently cluster related calorimeter cells and associate them with the particle type which initiated the shower. Clusters which can then be linked to reconstructed particles in the tracking devices will be removed from consideration, since their momenta will have been measured better by the tracking detectors. Only clusters unassociated with charged tracks will be used in the jet finding algorithm. In principle, these will be due only to photons and neutral hadrons. The photons will be well measured in an electromagnetic calorimeter. Reconstructing the energy deposits originating from long-lived neutral hadrons will be the most difficult task, and even though this class of particle represents a small fraction of the jet's composition, it represents an important part of the systematic uncertainty in the jet energy reconstruction.

CLUSTERING

In complex events and within jets, multiple particles will deposit energy in the same calorimeter cells and showers will overlap. Fine calorimeter segmentation and good clustering are essential to resolve such showers. Additionally, an intelligent cluster splitting and merging strategy is needed. Due to the fine segmentation and the high density of particles resulting from the collisions, many calorimeter cells are hit, so an efficient clustering algorithm is also essential. In this paper we present an object-oriented implementation of a fast, efficient, generic clustering algorithm to solve this problem.

The algorithm is based on clustering with local equivalence relations and requires only one pass through the data to establish the clusters. The simplest implementation uses a Nearest-Neighbor algorithm, but the relations can be generalized to larger neighborhoods. The framework can also be extended to arbitrary dimensions.

Cells, Neighborhoods and Clusters

The basic unit for clustering is a Cell. The Cell contains an Index by which it is referenced. For instance, a two-dimensional cell Cell2D could contain an Index2D composed of integer indices i and j to indicate row and column. The Cell also contains a value for the energy deposited. Note that Cells are not required to know anything about their relationship to other cells. The topology of a detector is encapsulated by a Neighborhood. Given an Index, a Neighborhood is responsible for returning a list of neighboring Indices. Since the dimensionality of the problem is encapsulated within the Index, the clustering algorithm can be written very simply and its extension to higher dimensions is automatic. The current implementation returns neighbors in a user-defined region which can be asymmetric in Index space, i.e. one can search for nearest-neighbors in one index space, but expand to next-nearest neighbors in others. When developing calorimeter designs with varying segmentation in the transverse (r-ϕ, or r-z) and longitudinal directions (layer depth) it will be essential to be able to easily investigate these different clustering procedures.

Clustering Algorithm

The clustering algorithm is based on local clustering with equivalence relations [1]. Instead of attempting to immediately and completely identify all the connections of a Cell under consideration, the idea is to make only the most certain associations. For instance, in a calorimeter one would most certainly wish to connect a Cell to its highest-energy neighbor. Repeating this procedure for each of the Cells in the calorimeter then defines the global clustering via the local equivalence relation routines. For efficiency, the list of Cells is first sorted by the Cell values (deposited energy). For each Cell in the list, one loops over all the neighboring Cells (provided by the Neighborhood) and establishes an association with its highest-valued neighbor. The connection is represented by a pair of pointers (or references) held by each Cell. The reference pointsTo in Cell i points to the Cell j which is its highest-energy neighbor. Similarly, the reference pointedTo in Cell j points to the Cell i for which it is the highest-energy neighbor. Since the list of Cells is ordered by energy, one can efficiently terminate the clustering loop when the Cell energy falls below a desired threshold. At the end of the process, linked lists of Cells comprise a Cluster. Note that isolated Cells point to themselves and thus form a single-Cell Cluster. A Cluster is a relatively lightweight object; it simply encapsulates a list of constituent Cells.

Cluster Fitting

The clustering algorithm is very efficient in resolving nearby clusters, but one still is faced with the task of splitting the energy shared between Clusters or merging the energy in nearby Clusters. Splitting the energy between neighboring Clusters requires knowledge of the shower shapes, which is reasonable for electromagnetic showers or muon traces, but is not as obvious for hadronic showers. We have only handled the electromagnetic case to date. A general non-linear multidimensional fitter has been written to allow n-dimensional Gaussian or exponential functions to be fit to identified Clusters. Each cluster is fit separately to establish initial estimates for the cluster parameters, then a global fit is performed on all nearby clusters. One can then subtract contributions from neighboring clusters or merge clusters if deemed appropriate. Clusters can also be incorrectly identified as separate, in the case of energy fluctuations within a shower, or correctly reconstructed as separate clusters but belonging to a single particle, especially in hadronic showers. In the former case, it is expected that a combined fit to neighboring clusters will identify cases of spurious local maxima, and that clusters identified as such will be subsumed into their parents. Separated clusters arising from charged hadron showers will be identified by their proximity to the extrapolated charged particle track reconstructed in the central detectors. We currently do not have a general strategy for merging separated clusters arising from neutral hadron showers. It is the immediate goal of this project to provide a framework for the development and evaluation of different clustering algorithms to resolve exactly this problem.

Ongoing Studies

The clustering and fitting code has been incorporated into the Linear Collider Detector software framework. Single Monte Carlo particles (electron, muon, pion and photon) sampled from the expected phase space to be encountered at a 500 GeV collider have been generated and the response of the proposed detectors has been simulated to create catalogs of shower shapes. Preliminary results for both cluster-finding efficiencies and cluster energy resolutions are quite promising and work is ongoing to develop not only the parametric representations of the electromagnetic shower shapes but also the criteria to be used to associate separated clusters of the hadronic showers. Once the full machinery is in place, actual studies devoted to the calorimeter detector design will be initiated. It is envisioned that a hyper-segmented calorimeter will be modeled and various realistic segmentations will be realized by ganging the Monte Carlo readout into appropriately sized calorimeter cells. In this way, multiple detector designs and reconstruction strategies can be run concurrently on the same events, thus minimizing the systematic uncertainties as well as the time and effort required to set up and run multiple scenarios.

ACKNOWLEDGMENTS

The author's work was supported in part by the U.S. Department of Energy under Contract DE-AC03-76SF00515.

REFERENCES

1. S. Youssef, Computer Physics Communications **45**, 423-426 (1987).