

Report on “Multiobjective Algorithm Parameter Optimization Using Multivariate Statistics”

Cecilia Chao Chen

This paper proposes a framework for tuning algorithm parameters in two levels. The first level, intra-experiment optimization, is to run simulations on a set of parameter values with minor changes in some experiment variables, then analyze the resulting values of multiple objective functions to discover their dependencies on the parameter, based on which an optimal parameter range is selected. In the second level, inter-experiment analysis, further investigation is done by repeating the intra-experiment optimization but each time with some major changes in experiment conditions, and the dependency of resulting optimal ranges on these major changes is studied and statistically modelled.

The first novelty of this paper is considering a large number of objective functions (typically expressed as FOMs – Figures of Merits), so the optimized parameter values achieve a generally satisfying performance. In most previous work only one or a few FOMs are considered, thus the optimal value for one FOM may result in a poor algorithm when evaluating with other FOMs. The algorithm discussed in this paper, *ART+blobs* for 3D EM (3D reconstruction in electron microscopy), has one free parameter λ . Certain values of λ achieve a high-contrast reconstruction but it is very noisy at the same time; if we focus on diminishing noise as much as possible the reconstruction would become very low in contrast. Neither is desirable for a general reconstruction tool. The methodology of determining λ in this paper, however, is able to find a compromise among FOMs of different purposes, by discovering their underlying relations and analyzing their dependencies on λ . The methodology for this is in fact the second contribution of this paper discussed next.

As reported in previous papers, the only way of parameter optimization is by experimenting: take a training FOM and a training data set, apply the algorithm with various parameter values and search for those values that optimize the training FOM for the training set. In a few papers where more than one FOM is considered, this routine is done for each FOM individually. This approach is unable to study the dependencies of various FOMs simultaneously, while this paper achieves this task using multivariate statistics. With a large group of FOM values from simulations on a representative 3D EM data set, ANOVA (analysis of variance) is carried out to study the dependency of each FOM on parameter λ : some insensitive FOMs are removed, while the remaining ones are hierarchi-

cally clustered based on their similar dependency trends. A single representative is then obtained for each cluster by PCA (principle component decomposition), and those λ values reflecting a reasonable performance in all clusters are chosen to form the optimal set. During the above process, analysis on the practical meaning of each cluster and its dependency trend reveal the effect of λ on different performance goals, thus providing not only a generally optimal λ , but also a guideline of choosing λ according to specific FOMs.

The third contribution of this paper is the inter-experiment analysis. Consider the factors affecting parameter choice and algorithm performance: some minor changes, such as small difference in particle size, different level of noise and angular distribution of projections in 3D EM, may cause different optimal λ but in a minor and random manner; some major changes, such as number of projections, sampling rate and noise nature in 3D EM, may influent the optimal range of λ to a rather significant extent. ANOVA is used to study whether the optimal range depends on some major changes, and if it does, how much the dependency is. Furthermore, a statistical model of the dependency is established and determined with nonlinear regression analysis. This model can then be used to estimate an optimal parameter range when the algorithm is applied to similar experiment conditions in the future.

In summery, the framework of multi-objective optimization achieves reasonable performance for various FOMs, and analyzes the influence of experiment conditions on the optimal parameter range. The methodology is strongly based on multivariate statistics, therefore is expected to be remarkably useful for algorithms in different areas. On the other hand, due to its strong statistic nature, two issues need great concern: choosing the representative data set and determining the interval of parameter values in experiments. A good selection of data set is essential to the representativeness and applicability of the resulting optimal parameter values; The interval of parameter candidates, λ as in this paper, affects how well each FOM's (thus each FOM cluster's) dependency on λ is discovered, and how reasonable the statistical model for the dependency on experiment conditions could be. An easy answer would be having a large set of data and a small interval as possible, but in many applications the complexity of the algorithm makes it expensive to repeat experiments for many times. Therefore these two factors should be carefully chosen. Once a reasonable set of experiments and analysis are done, the results would be very useful for those experiment conditions and objective functions similar to what have been studied, since the major effects of the parameter have been captured, and an optimal range can be easily computed. The last thing to mention is that, many details in this framework could be modified to adapt to specific purposes, such as clustering method, cluster number choosing, optimal range determining, and especially, the modelling of dependency on major changes. With a prior knowledge and special preferences, this framework would provide us with more trustable results.