ORBIT CORRECTION METHODS – BASIC FORMULATION, CURRENT APPLICATION AT JEFFERSON LAB, AND FUTURE POSSIBILITIES

Yu-Chiu Chao
Thomas Jefferson National Accelerator Facility, Newport News, VA 23606 USA

Abstract

A. Orbit Correction System Optimization: Recipes for optimizing an orbit correction system configuration at the design level are presented. Linear algebraic tools are applied to various flavors of response matrices to uniformly control unobservability, uncorrectability, and response matrix singularity. Application at Jefferson Lab is discussed. *B. Orbit Correction at Jefferson Lab*: Unique challenges posed by orbit correction, as well as algorithms and tools developed at the CEBAF accelerator at Jefferson Lab are discussed. *C. Orbit Interpretation and Virtual Monitors*: A new approach to developing an orbit correction package with software structural, algorithmic and operational advantages is introduced. It consists of an orbit interpretation module, a virtual monitor module, and a generic steering engine. Mathematical formulation, algorithms prototyped and tested on simulated and real data, and future possibilities are discussed.

1 INTRODUCTION

Orbit correction has been among the most studied problems of accelerator control. Algorithms have been developed at various laboratories to meet specific demands. Some of these algorithms, such as MICADO [1] or SVD [2], have found much wider applicability than was originally envisioned, and are approaching the status of universal steering engines through their many reincarnations.

The main purpose of this report therefore is not to introduce one more steering algorithm, but rather to present a formalized approach to orbit error and correction, with steering engines being one (important) link in the larger process of accelerator system design, operation and improvement. From this approach we are able to formulate orbit correction issues in an analytic framework, and develop quantitative design criteria, recipes for system optimization, and tools for visualizing and controlling various types of errors under a single unified scheme, even before orbit correction is applied.

The unique design and physical constraints of the CEBAF accelerator at Jefferson Lab imposed special demands on orbit correction, which inspired most of the studies presented here. Application of these tools and general experience on orbit correction at CEBAF will be discussed throughout this report. Despite their somewhat parochial origin, these tools were developed with the most generic orbit correction system in mind, and should be universally applicable regardless of the specifics of a given system.

Throughout this report extensive use will be made of response matrices, which characterize not only the linear behavior of the orbit correction system, but also, when generalized, the error-induced orbits and the unobserved effects of orbit correction. Complete knowledge of these generalized response matrices affords quantitative predictions on the global performance of an orbit correction system, and ability to control errors at a higher level. The advantage of response matrices over more intuitive methods, such as betatron phase counting, in analyzing orbit correction problems should also be noted, in that the former can always give unambiguous answers in otherwise ambiguous situations.

Orbit correction can fail for a number of reasons. We can nonetheless place the blame on either of the two fundamental causes: design flaw (static) and run-time system breakdown (dynamic). These are listed in Table 1.1[1]. A good orbit correction algorithm should successfully handle problems related to response matrix degeneracy and input error, and provide insight into dynamically generated uncorrectability. It is however advisable, and usually unavoidable, to address static unobservability, response matrix singularity, and uncorrectability by re-configuring the orbit correction system.

Table 1.1
Problems Encountered in Orbit Correction (Anywhere)

PROBLEM	SYMPTOM	SOURCE	EXAMPLE
Response matrix degeneracy	Excessive correction Unobserved orbit error	Static Dynamic	Redundant correctors Missing monitors
Fundamental unobservability	Over-sensitivity Poor reproducibility	Static Dynamic	BPM deficit by design Missing monitors
Fundamental uncorrectability	Large residual orbit	Static Dynamic	Corrector deficit by design Corrector Limit Misalignment / Injection error Unaccounted kick
Error in input data	Undetectable orbit error	Dynamic	Bad BPM calibration
Model error	Breakdown of correction Failure to converge	Dynamic	Quadrupole gradient error Multipole components in dipole

The very act of steering impacts parameters other than orbit at the beam position monitors (BPM). These parameters are important for machine performance in general, and crucial for the successful operation of CEBAF due to its unique operational requirements. Table 1.2 shows a list of such parameters relevant to CEBAF, the area of relevance, their impact on machine performance, and the agent responsible for coupling them to the monitored orbit. The number of recirculations (5) in the linacs is small enough to make simultaneous multiple pass steering possible, and large enough to make it almost a necessity. This is also included in Table 1.2.

Table 1.2
Generalized Orbit Correction Scenarios (at CEBAF)

OBJECTIVE	AREA	IMPACT	COUPLING
Energy calibration	Arc	Energy/path length/ dipole string/setup	Dispersion & energy feedback
Angle control	Re-injection Spreader/Recombiner	Baseline setup	Betatron propagation
Path length control	Spreader Recombiner	RF synchronization	Betatron propagation
Dispersion control	Spreader/Recombiner	Energy stability	Chromaticity
Orbit at unmonitored locations	Septum magnets Between dipoles	Baseline setup	Betatron propagation
Multiple pass orbit	Recirculation linac Spreader/Recombiner	Baseline setup	Common steering elements

In Section 2 a complete recipe based on extended response matrices is given for eliminating static configuration flaws in an orbit correction system with minimal set of elements, thus optimizing the balance between performance and economy. Application at CEBAF is discussed. In Section 3 we describe the special challenges posed by orbit correction at CEBAF, and the tools developed to meet them. In Section 4 a new algorithmic scheme is introduced with a more global approach, taking into account the underlying orbit, sources of error, and generalized constraints. Out of this approach a self-contained orbit interpretation and control program emerges with software structural, algorithmic, and operational advantages. The errors of Table 1.1 are revisited under this formalized scheme.

[1]Empirical model and adaptive algorithms are two solutions to model errors, which will not be covered here.

2 ORBIT CORRECTION SYSTEM OPTIMIZATION

As stated in Section 1, static steering problems arising from configuration flaws are best addressed by system re-configuration to ensure that the orbit correction system performs at a desired level everywhere. In the following we present a proven set of recipes [3] aimed at configuring orbit correction system with optimal balance between performance and economy. Through this program the three static problems in Table 1.1, unobservability, response matrix singularity, and uncorrectability, are controlled uniformly to within tolerances defined by operational needs.

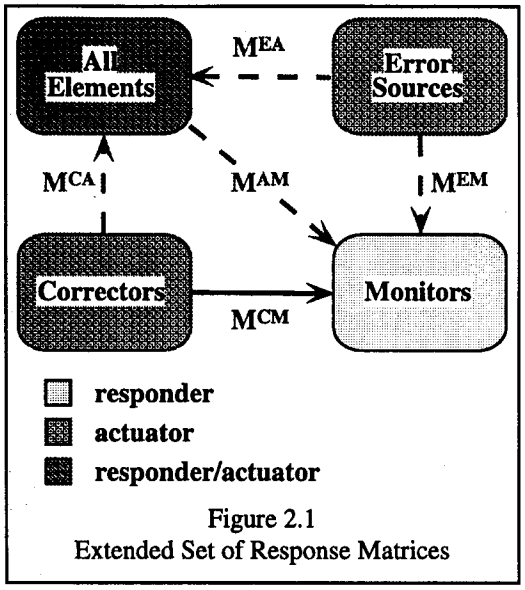

Figure 2.1
Extended Set of Response Matrices

The essence of this program lies in extending the scope of the response matrix beyond its description of the real actuators (correctors) and responders (BPMs), to that of the "virtual" ones. A set of linear algebraic tools, SVD being among the most useful, can be readily applied to such an extension, yielding insight into the global behavior of the orbit correction system. Fig. 2.1 shows, in addition to the real response matrix M^{CM}, the extended response matrices, in dashed lines, constructed out of "virtual" actuators and responders. As these matrices will recur throughout the report, they deserve more explanation[2].

2.1 Extension to more general response matrices

2.1.1 Error-to-monitor response matrix M^{EM}

The error-to-monitor response matrix M^{EM} summarizes the disturbance in any of the beam coordinates at all monitors by all potential physical errors. The latter includes injection errors, magnetic field errors, misalignments etc., and the actual matrix elements consist of optical transfer elements M_{11}, M_{12}, and M_{16} between error locations and monitors. In constructing M^{EM} one must identify all potential sources of errors the orbit correction system is designed to correct. To make M^{EM} more realistic, design tolerances on magnet and alignment can be incorporated by scaling individual columns of M^{EM} to reflect the design characteristic.

2.1.2 Error-to-all-location response matrix M^{EA}

The error-to-all-location response matrix M^{EA} summarizes the orbit disturbance at all representative locations caused by all physical errors described above. These representative locations, not tied to any physical elements, should effect coverage of the beam line dense enough to capture all potential orbit extremes. They will be collectively denoted by a set C_A, which typically consists of all the electro-magnetic element locations, critical ends of drifts, and any user selected location of interest.

2.1.3 All-location-to-monitor response matrix M^{AM}

The all-location-to-monitor response matrix M^{AM} summarizes the orbit disturbance at all monitors caused by coordinate errors at all representative locations in the set C_A.

2.1.4 Corrector-to-all-location response matrix M^{CA}

The corrector-to-all-location response matrix M^{CA} summarizes the orbit disturbance at all representative locations caused by all correctors.

---

[2] For simplicity we describe everything in the x-plane only with conventional coordinate assignments of 1, 2, 5, 6 for position, angle, path length and momentum offset.

2.2 The optimization program

After establishing extended response matrices for a given hardware and optics configuration, we can look into the global performance of the orbit correction system by applying various linear algebraic tools to these matrices. The optimization program described in this section provides efficient, quantitative and unambiguous answers that may elude intuitive inspection of the optical lattice or numerical simulation. In addition a set of recipes establishes the path to reaching desired tolerances on unobservability, response matrix singularity, and uncorrectability.

The program starts with a configuration of orbit correction system, namely, a set of correctors and BPMs, out of which all extended response matrices are constructed. It proceeds to check for the following configuration flaws and remove them iteratively in exactly the order given below[3,4].

2.2.1 Monitor deficiency → Fundamental unobservability

This is the situation where well behaved orbits at all monitors cannot guarantee the same everywhere. In other words, blind spots exist making potentially harmful orbit at some location unknowable.

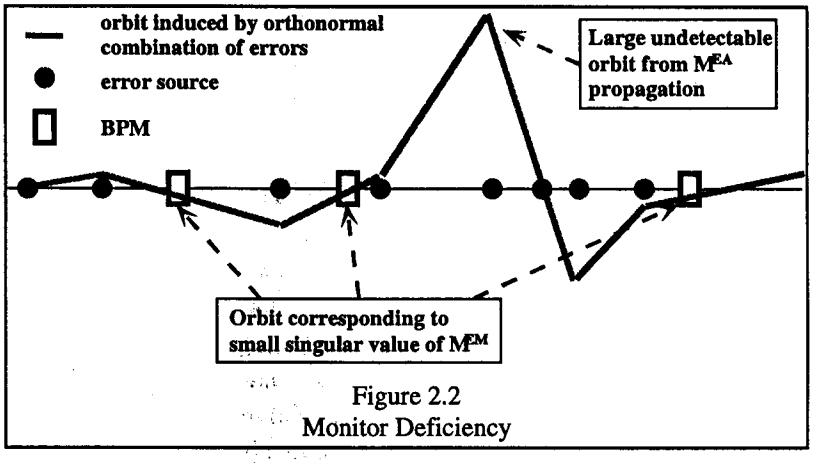

Figure 2.2
Monitor Deficiency

First the generalized error-to-monitor response matrix M^{EM} is properly scaled to reflect design magnet and alignment tolerances. Unobservability is indicated by the presence of null space vectors or very small singular values of M^{EM}. In both cases the system fails to meet a numerical criterion defined through operational requirements. The matrix M^{EM} is singular value decomposed (SVD) into ortho-normal combinations in the corrector space. Combinations corresponding to singular values short of the numerical criterion are identified, indicating unobservable error effects. The error-to-all-location response matrix M^{EA} is then applied to these combinations to get error-induced orbits at all relevant locations. The largest element in each of the orbit vectors is identified and, if this number exceeds a second numerical criterion for acceptable unobserved orbit, a new monitor is added at this element or its vicinity. The procedure is iterated until M^{EM} no longer has null space vectors or singular values smaller than the criterion. A typical step leading to added monitor is shown in Fig. 2.2.

2.2.2 Monitor redundancy → Hardware excess & unjustified corrector requirement

The reason that we need to worry at all about too many monitors is that once the minimally necessary set of monitors is achieved, adding more would not improve observability, and may even place unjustified constraint on the correctors in the following step.

The program starts with the projection operator Π^{EM} [3] associated with M^{EM}, which divides any orbit vector into parts inside and outside the subspace spanned by the columns of M^{EM}. Π^{EM} is applied to all unit vectors in the orbit space representing unit offset at each BPM. The length of each resulting vector, representing coupling between errors and monitors, is calculated. All BPMs with

---

[3] Conceptual outline rather than detailed procedure will be given here as the latter can be found in Ref. [3].

[4] It should be noted that general accelerator design respects the following numerology: $N_E > N_M \geq N_C$, where N_E, N_M and N_C are the total number of potential errors, monitors and correctors respectively. Thus the matrix M^{EM} has more columns than rows and the opposite is true for M^{CM}. The optimization program is most relevant when this is the case, and if not, will likely leave the system in this state at the end.

length smaller than a numerical criterion are deleted as they do not reflect enough error of interest.

The program then moves on to orthogonality checks, which is done with the Gram determinant [3] to ensure absence of redundancy in the observable orbit. If a numerical criterion is not met, iterative elimination of monitors is done with SVD on M^{EM} and identification of the dominant monitor in the resulting monitor combination with the smallest singular value. This is repeated until M^{EM} passes the Gram-determinant test.

2.2.3 Corrector deficiency → Fundamental uncorrectability

This is the most basic requirement of any orbit correction system, namely, there must be enough correcting power to counteract errors, whether injection, electromagnetic kick, or misalignment.

The program performs SVD on M^{EM} to obtain ortho-normal combinations of error-induced orbit vectors. Or, if there is enough confidence in the last two tests on monitor behavior, all unit vectors in the orbit space can be used instead with proper scaling reflecting design error tolerances. The projection operator Π^{CM} associated with the real response matrix M^{CM}, and the pseudo-inverse of M^{CM} [3], are then applied to all orbit vectors to obtain the uncorrectable fraction and the required corrector strength for each orbit vector. Both outcomes are subjected to numerical criteria to identify corrector deficiency. Iterative addition of correctors is achieved through comparing projection of column vectors of the all-location-to-monitor response matrix M^{AM} and the "residual" orbit vector derived from the deficiency test above. Iteration ends when both numerical criteria are met.

2.2.4 Corrector redundancy → Hardware excess & response matrix singularity

This is the cause of excessive corrector strengths and unobservable orbit excursion resulting from orbit correction. If economy of hardware is not a concern or surplus correctors are needed for special purposes, this program can be skipped and the response matrix singularity left to be handled by smart steering algorithms. For example, the virtual monitor algorithm described in Section 4 addresses this problem with algorithmic and operational advantages.

The program performs SVD on M^{CM} and evaluates its condition number, a measure of the evenness in the corrector effect distribution. It also calculates the Gram determinant of M^{CM} to determine the orthogonality of the corrector effects. Both are compared to numerical criteria. If either criterion is not met, the index of the largest element in the SVD generated ortho-normal corrector combination corresponding to the smallest singular value is identified. This index points to the corrector to be removed unless it corresponds to a deliberately retained corrector, in which case the corrector corresponding to the next largest element is removed. This is repeated until both criteria are met. The criteria for corrector non-deficiency used in the last step should be monitored at each iteration to prevent over-reduction.

2.3 Application at CEBAF

The optimization program was used during commissioning of the CEBAF accelerator to provide a quantitative guidance on orbit correction effectiveness. Orbit correction system in every section was subjected to the tests described to ensure same level of performance. Areas where steering difficulties were encountered were found to be the same areas that stood out significantly in the tests. The recipes described above were then employed to re-configure the orbit correction system until it passes the tests. The improvements have been corroborated by improved steering reproducibility, corrector strength and orbit excursion, which translate into operational and even optical gains.

An additional advantage of formally optimizing the orbit correction system was realized when steering algorithm was developed for CEBAF, demanding various exception handling measures built into the algorithm. We were able to refer to the baseline configuration for estimates on effectiveness of such measures, knowing in the first place that the former was free of static problems.

2.3.1 Monitor Deficiency

It was realized from the monitor deficiency test of *2.2.1* that undetectable error-induced orbits inside all 5 passes of the East Extraction Region were 5 times larger than anywhere else in CEBAF. This was supported by high steering sensitivity and poor corrector reproducibility in this region regardless of method of steering. According to simulation this undetectable orbit could sample multipole components in nearby dipoles and cause emittance distortion of up to 10%. The optimization program identified 5 new monitor locations to bring undetectable orbit level in line with the rest of the machine. Steering efficiency and corrector reproducibility have improved to the same level as the rest of the machine since these monitors were installed.

2.3.2 Corrector Redundancy

Excessive correction in lower arcs and poor reproducibility in spreaders and recombiners during machine setup at CEBAF suggested excessive coverage of beam line by correctors. The corrector redundancy test of *2.2.4* was applied to the entire accelerator and correctly identified the most offending correctors in lower arcs, with singularity index 20 times greater than anywhere else. It also established a prioritized sequence of corrector removal in spreaders and recombiners. Corrector deficiency criteria were monitored at each step to prevent over-reduction. The machine has been operating with this reduced corrector set. Neither previous steering problem nor compromise in orbit correctability has been observed.

3 ORBIT CORRECTION AT JEFFERSON LAB

Jefferson Lab operates its CEBAF accelerator, with which it is often synonymous, as a nuclear physics research facility currently delivering CW electron beam to three fixed-target experiments with energy up to 5 GeV. CEBAF consists of injector, multi-pass linacs, re-circulating arcs, beam separation (spreader) and recombination (recombiner) structures, and extraction lines to experiments. These are shown in Fig. 3.1. The complex trajectory manipulation, often simultaneous in multiple passes, the need to satisfy multiple beam and optical constraints in both transverse and longitudinal dimensions, the deviation from smooth, periodic FODO lattice in many areas, and the space limitation on instrumentation all guarantee a rich environment for steering challenges. These challenges, as well as existing and proposed solutions, are discussed in the following subsections.

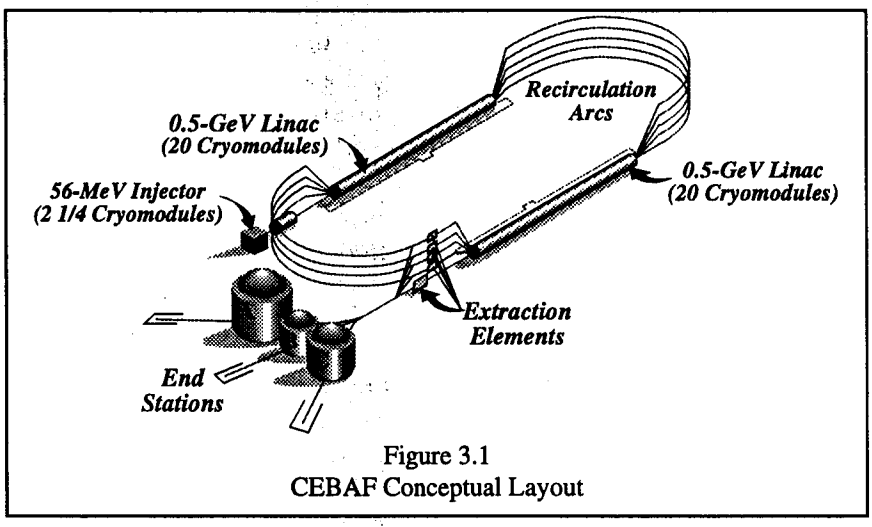

Figure 3.1
CEBAF Conceptual Layout

3.1 Steering engine

A locally developed algorithm, PROSAC (Projective RMS Orbit Subtraction And Reduction), is used at CEBAF as the generic steering engine with a strong emphasis on fully exploiting hard corrector limits while strictly conforming to them[5]. This turns out to be a valuable feature in many cases.

[5] An alternative method of eliminating corrector combinations through SVD when corrector limit is reached is intrinsically pathological. Since SVD only deals with orthogonality of the response matrix, it can misinterpret large corrector values caused by fundamental uncorrectability as singularity-induced, and proceed to over-eliminate correctors. An algorithm for correctly eliminating correctors under SVD has been conceived at CEBAF but

The principle of PROSAC is very simple. All correctors are mapped by the response matrix into "effective orbit" vectors, which are compared in turn to the real orbit vector and the one with the largest projection on the real orbit, either normalized or un-normalized, is used to truncate the orbit vector up to the corrector limit. This process is iterated until a user-defined reduction target is met. Corrector prioritization is also carried out to let the user apply the most effective subset of the correction while achieving most of the correction goal.

Accompanying PROSAC are exception-handling algorithms for either data pre-processing to eliminate errors or guarding against operational problems. These are discussed in the following.

3.2 Solution to dynamic problems

Dynamically occurring steering problems at CEBAF that require algorithmic exception handling are a subset of Table 1.1. We will describe them below.

3.2.1 Dynamic response matrix singularity and unobservability

With the absence of static monitor deficiency and corrector redundancy guaranteed, response matrix singularity happens only when some BPM's become unavailable at execution time. Two algorithms were used to prevent adverse effects of excessive correction due to near-singularity.

- Corrector elimination: SVD is performed on the response matrix and a procedure similar to that described in Section 2.2.4 is executed.
- Additional orbit constraint: Trajectory fitting is performed on the orbit and corrector data. The projected orbit at missing BPM's is added to the real orbit before steering.

These methods can in principle be either too heavy-handed or misrepresenting reality if too many BPM's are missing[6]. In practice, however, they have kept the steering process from diverging.

It should be noted that the real cause of the difficulty is that, with missing BPM's, we also get fundamental unobservability. When too many BPM's are missing, the system configuration is no longer adequate for orbit correction.

3.2.2 Error in input data

Systematic offsets in BPM readings present major impediment to successful orbit correction. At CEBAF the procedure of "quadrupole centering" is performed on BPM's at critical locations to mitigate this problem. This is done by varying the strength of the quadrupole next to the BPM in question while changing the beam position at the same location. Beam is considered "centered" when downstream orbit oscillation induced by quadrupole variation is smaller than a specific tolerance. BPM database is then updated to reflect this center. Currently this procedure is done manually at strategic areas to provide zero-th order information on the machine baseline.

BPM offset error at CEBAF has partly inspired the orbit interpretation algorithm, to be discussed in Section 4, with the aim of separating fundamental uncorrectability from monitor input error, and deducing the underlying orbit for correction. This algorithm has been successfully applied to real data at CEBAF to identify BPM's with offset errors and to resolve the underlying orbit and dipole kicks in linacs and spreaders.

Finally built into PROSAC is a crude version of the orbit interpretation algorithm which aims at identifying extreme BPM offsets. This is invoked as a user option.

---

not tested on line. The optimal approach to the corrector limit problem, correctly handling both fundamental uncorrectability and response matrix singularity, would be using PROSAC on orbit generated by the virtual monitor algorithm, to be discussed in Section 4.

[6] Again, the correct way to address this problem is to use virtual monitors or equivalent algorithms, if not too many BPM's are missing.

3.2.3 *Dynamic uncorrectability*

This is usually not a cause of concern unless uncorrectable orbit indicates anomalous injection or large unaccounted disturbance. The orbit interpretation algorithm discussed in Section 4 has also been motivated by attempts to identify this effect. Currently no exception handling is built into PROSAC to address this problem, other than indirect inference from extreme apparent BPM offsets.

3.3 Special steering scenarios

A summary was given in Table 1.2 of special steering scenarios encountered at CEBAF. We will describe below methods developed or planned to accomplish these tasks.

3.3.1 *Energy calibration*

The lowest energy arcs in CEBAF are used as spectrometers to both calibrate and stabilize energy out of the linacs. Energy stabilization is possible due to large dispersion in the arcs exploited by the energy feedback system. But before feedback can be invoked linac energy has to be matched to the arc dipole with a "standard" arc orbit pattern established. PROSAC is employed to perform the task of simultaneously establishing the standard orbit and matching the energy. This is done by including the momentum offset dp/p as an extra corrector with associated M_{16} as response matrix elements, and forcing the averaged physical corrector strength to that needed only for counteracting earth field while allowing individual variations to correct local orbit. This procedure decouples the non-dispersive orbit from the dispersive orbit and corrects both in a single step. The same algorithm has been used for higher energy arcs where linac energy can no longer be changed, but simultaneous correction of orbit and path length, or orbit and main dipole strength, may be desired.

3.3.2 *Angle control and beam threading*

With many junctions between functional modules of CEBAF, clean injection from section to section is an important issue. Currently PROSAC can be configured to perform injection optimization into multiple pass linacs where local steering is not possible. This is done with orthogonal correctors to cover injection phase space, with orbit in the downstream section providing target of correction.

For other sections PROSAC provides the "zero angle" option, which freezes the strengths of all correctors between the last two BPM's at zero before applying correction. If the remaining correctors are not driven to limits, and if the optical transfer between the last two BPM's is not close to point-to-point imaging, then the outgoing beam angle should be close to zero. This technique has also been used for beam threading with PROSAC when the arcs were first commissioned.

Real and absolute angle control, nonetheless, is possible only with the introduction of virtual monitors, to be discussed in Section 4.

3.3.3 *Steering with common dipoles*

In the spreaders and recombiners of CEBAF main vertical bending dipoles have significant effects on the orbit. Due to the machine setup sequence dipoles common to multiple passes can couple lower pass injection errors to higher passes. Off-line orbit interpretation has been performed to disentangle multiple pass effects with reasonable results. During commissioning a prototype program has also been developed to use, among other correctors, the common dipoles to simultaneously fix injection and correct local orbit for multiple-pass spreaders and recombiners.

With CEBAF in its production phase, the need to adjust individual dipoles has decreased significantly. However in order to respond to flexible energy requirements from the experiments, it is planned for the next phase of PROSAC to include some of these dipoles as orbit correction elements for more efficient baseline setup.

3.3.4 *Path length control, dispersion control, and orbit at un-monitored locations*

Absolute control of these parameters are not possible without the virtual monitor algorithm, which is

not implemented at CEBAF currently. They are planned for the next phase of PROSAC.

3.4 Simultaneous multiple pass steering in the linacs

In early 1998, large and persistent orbit patterns in all 5 passes inside the CEBAF linacs were seen to develop, defying pass-by-pass correction. This was blamed on possible misalignment and unaccounted disturbances. Significant difference in betatron phase advances between different passes and absence of correctors exactly coinciding with all potential errors left higher pass orbits at the mercy of first pass corrections. This was exacerbated by unknown systematic offsets in the BPM's. Effort was first made to determine the underlying errors, including injection, sources of kicks, and monitor offsets. This highly constrained analysis yielded very reliable estimates on monitor error and underlying orbit. It was realized from simulation that using all the correctors inside the linac, which affects all passes differentially, as well as injection fixes from individual upstream recombiners, we could reduce the orbit in all passes significantly. This was done in the South Linac, using PROSAC as the steering engine[7]. Fig. 3.2 shows the outcome with all 5 pass orbits displayed in tandem for each plane. The solid line in x-plane is an order of magnitude smaller in RMS than the dashed line[8]. A total of 12 horizontal and 13 vertical correctors in the linac and 10 (2×5) correctors in each plane in the recombiners were used to achieve this orbit reduction at 135 locations (27 BPM ×5 passes) in each plane [5].

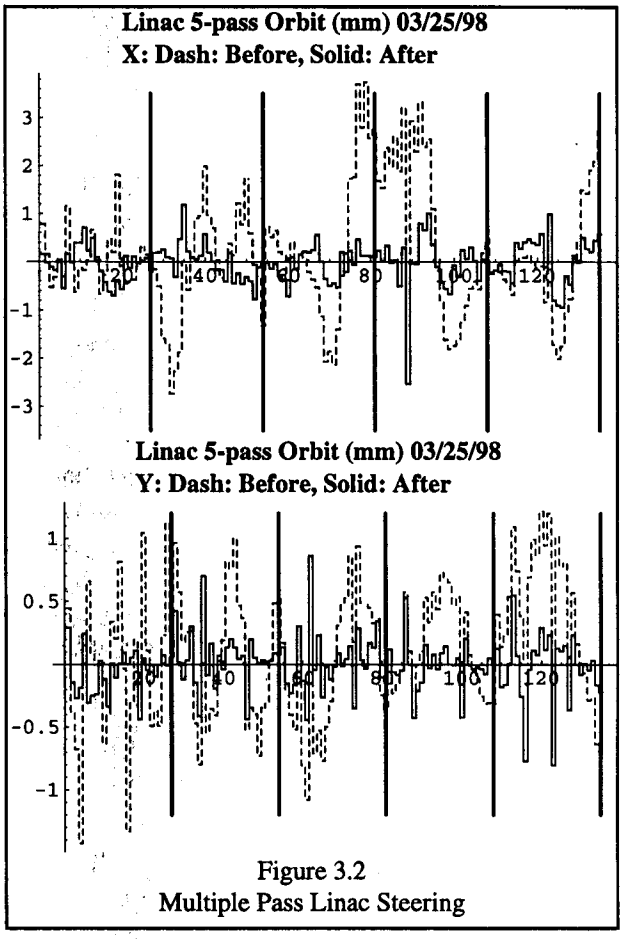

Figure 3.2
Multiple Pass Linac Steering

4 ORBIT INTERPRETATION AND VIRTUAL MONITORS

In earlier sections we discussed exception handling methods for minimizing dynamically occurring steering errors, as well as for controlling generalized beam and optics parameters. It was noted that these methods are less than perfect mainly because of ignorance about the underlying errors. This is what prompted a study into ways of approaching the orbit correction problem with a more comprehensive and consistent formulation such that all types of errors and their effects on steering are accounted for under a single unified scheme. From this formulation we can gain insight into the underlying problem, disentangle errors with overlapping signatures, reconcile between conflicting steering objectives, and create arbitrary steering scenarios with optimal exception handling ensured.

A conventional way of developing orbit correction program is shown in Fig. 4.1 where steering constraints at CEBAF are used as an example. In this approach a steering algorithm, for example SVD, is coded to solve the most generic steering problem. More Sophisticated features, such as needed for exception handling and generalized constraints, are then added as operational need develops. Complexity multiplies as the program is modified to accommodate these features. In the end one has a steering program which can meet a unique set of operational requirements.

---

[7] Some correctors were driven to their design limit in the process, justifying the choice of PROSAC.
[8] Most solid spikes correspond to malfunctioning BPM's.

The shortcoming of this approach is obvious. From the software viewpoint, incremental and ad hoc feature enhancements undermine modularity and flexibility, making it harder to incorporate alternative error handling features and steering algorithms. From the machine operational viewpoint, information inside the program is inaccessible to other interested clients such as magnet or alignment. There are also algorithmic limitations to orbit correction with this approach to be discussed later.

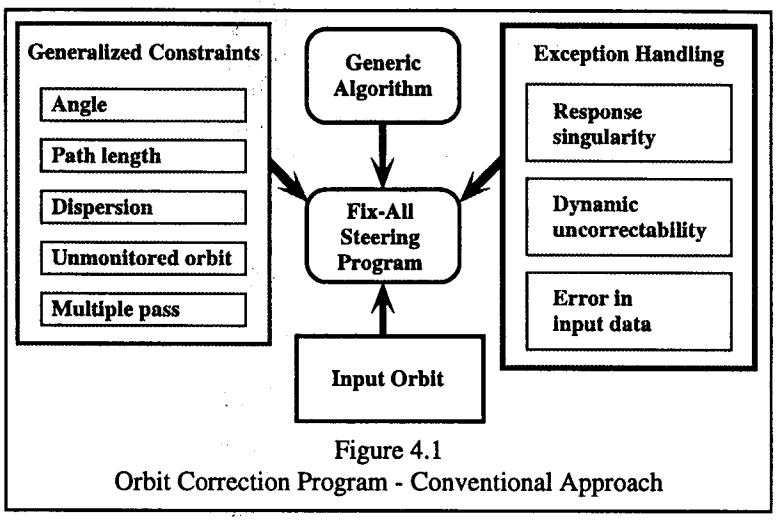

Figure 4.1
Orbit Correction Program - Conventional Approach

We can on the other hand widen the scope and follow the alternative approach of Fig. 4.2[9]. The orbit, hardware, and model data are input to an orbit interpretation module, which interprets the underlying error and orbit by separating contributions from fundamental uncorrectability and input monitor error. The algorithm used for this analysis can be interchangeable since it responds to a well defined, fixed set of inputs. The output of this module is available to other clients interested in the performance of magnet, alignment, or diagnostics, as well as to the second component of this package: the virtual monitor module. The virtual monitor module does not perform any analysis, but upon input of the interpreted orbit[10], manipulates all generalized steering constraints, and most importantly, singularity control, into a set of "virtual" orbits and response matrices. The virtual objects are indistinguishable from the real ones for the steering engine. When they are input to the steering engine, the outcome automatically satisfies all constraints on generalized coordinates and response matrix singularity. It is also less prone to input monitor error due to screening by the orbit interpretation algorithm. The third component of this package is the generic, interchangeable steering engine.

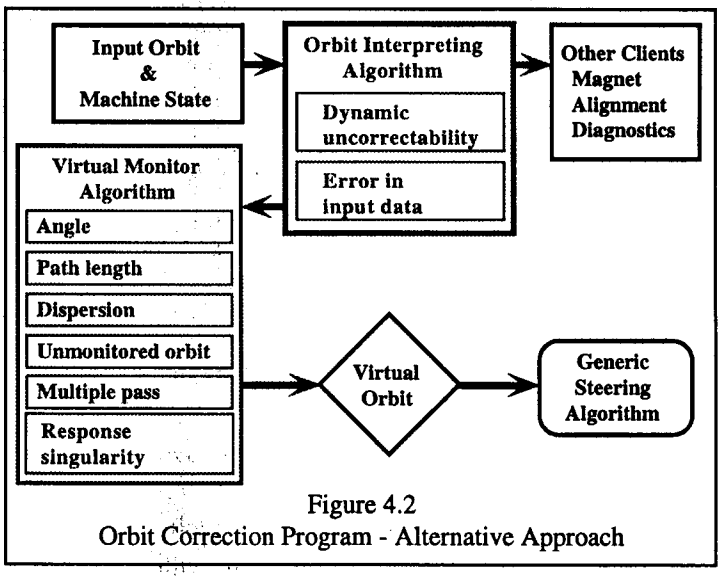

Figure 4.2
Orbit Correction Program - Alternative Approach

This approach exhibits multiple advantages over the conventional one, as elaborated below.

- Software structural advantage
  - All feature enhancements are done through interchanging orbit interpretation algorithm or steering algorithm, with well-defined and fixed inputs.
  - Feature enhancements will never touch the virtual monitor module, which executes a fixed procedure with all options anticipated.
  - Important information from orbit interpretation is available to other clients.

[9] Notice the migration of various components from Figure 4.1 to Figure 4.2.

[10] This will be given at all monitored and un-monitored locations and in all monitored and un-monitored coordinates. For example, the interpreted orbit can contain information about path length inside a dipole.

- Being an independent module as opposed to ad hoc add-on features, the orbit interpretation algorithm can be developed in a much more compact and consistent manner.

▪ Algorithmic Advantage

- Steering outcome inherits optimal screening of input monitor error by orbit interpretation.
- One can set absolute targets for generalized constraints, as opposed to actuating only increments without orbit interpretation.
- Controlling singularity through virtual monitors is superior to corrector elimination or correction strength minimization which, not fully exploiting the response matrix, can be too heavy handed.
- Singularity control through virtual monitors automatically takes into account constraints on generalized coordinates.

▪ Operational Advantage

- User can see the underlying picture, important in diagnosing anomalous cases.
- User can satisfy different steering objectives simultaneously, avoiding unnecessary iterations.
- User can create on-the-fly steering scenarios more easily and confidently.

As we will see below, the orbit interpretation and virtual monitor modules require minor modification to an existing steering algorithm. The main enhancement comes from establishing generalized response matrices of Section 2, which are fixed entities in the database once established.

4.1 Orbit interpretation

Let us begin by asking what makes up an observed orbit at a monitor located at point p. Staying within the x-plane for simplicity, we have

$$\begin{aligned} O_1^p &= \sum_k M_{12}^{kp} \cdot C_2^k && \text{corrector kicks} \\ &+ \sum_j M_{1j}^{0p} \cdot \delta x_j^0 + \sum_a \sum_j M_{1j}^{ap} \cdot \delta x_j^a && \text{injection \& misalignment errors} \\ &+ \Delta_1^p, && \text{monitor error} \end{aligned} \tag{4.1}$$

where we decomposed the x-orbit O_1 at p into contributions from correctors C at locations k, injection errors δx^0, misalignment errors δx at locations a, and monitor error Δ. Misalignment error stands for all field and geometry errors in the beam line that can change any beam coordinate. Since injection error satisfies this description, we can classify it under misalignment and rewrite Eq. (4.1) as

$$\begin{aligned} O_1^p - \sum_k M_{12}^{kp} \cdot C_2^k &= \sum_{a+0} \sum_j M_{1j}^{ap} \cdot \delta x_j^a + \Delta_1^p \,, \\ T &= K + \Delta \end{aligned} \tag{4.2}$$

where we have on the left hand side the "naked orbit" T, consisting of all known measurements, and on the right hand side all the unknown errors, including the generalized misalignment K, and monitor error Δ. If the naked orbit T is blindly input to any steering engine, the best outcome one can expect is determined by the projection operator associated with the response matrix M^{CM} given by

$$\begin{aligned} E &= \Pi^{\perp}_{\mathrm{M_{CM}}} \bullet K - \Pi^{\parallel}_{\mathrm{M_{CM}}} \bullet \Delta \,, \\ \Pi^{\parallel}_{\mathrm{M_{CM}}} &= \mathrm{M_{CM}} \bullet \left(\mathrm{M^T_{CM}} \bullet \mathrm{M_{CM}}\right)^{-1} \bullet \mathrm{M^T_{CM}} \,, \\ \Pi^{\perp}_{\mathrm{M_{CM}}} &= \mathrm{I} - \Pi^{\parallel}_{\mathrm{M_{CM}}} \end{aligned} \tag{4.3}$$

where E represents the residual orbit. The projection operators $\Pi^{\perp}$ and $\Pi^{\parallel}$ divide the orbit space into the part outside the subspace spanned by the column vectors of M^{CM} and that inside, or, orbit uncorrectable and correctable by M^{CM}. The residual E stands for the real residual orbit after correction even if the apparent residual orbit may be different due to monitor errors. Eq. (4.3) shows that K and Δ have opposite effects on the residual orbit, which consists of all misalignment errors uncorrectable

by M^{CM}, and all monitor errors correctable by M^{CM}. The former is fundamentally uncorrectable, but the latter is spurious, disguised as alignment errors to compromise steering effectiveness. The ultimate goal of orbit interpretation is to disentangle these contributions, with their mostly distinctive signatures, to achieve optimal steering[11]. To do this, we resort to the error-to-monitor response matrix M^{EM} of Section 2 and perform the analysis discussed in the sections below.

4.1.1 Using alignment errors as virtual correctors

Each column of M^{EM} can be regarded as the effect on the orbit by a "virtual corrector" corresponding to the misalignment error. A generic steering algorithm can be applied, using M^{EM} as the response matrix, to the orbit T of Eq. (4.2) to get

$$\begin{aligned} T &= \Pi^{\parallel}_{M_{EM}} \bullet (K+\Delta) + \Pi^{\perp}_{M_{EM}} \bullet (K+\Delta) \\ &= T_K + T_\Delta . \end{aligned} \tag{4.4}$$

We can then interpret T_K as the real orbit generated by misalignment errors, to be used for orbit correction, and discard T_Δ as spurious monitor error. In practice only a subset of M^{EM} can be used in Eq. (4.4) since the problem is highly under-constrained. Let U^0 be a sub-matrix of M^{EM} containing part of the column vectors of M^{EM}, substituting U^0 for M^{EM} in Eq. (4.4) and applying orbit correction in the fashion of Eq. (4.3) on the interpreted orbit T_K results in a new residual orbit

$$E = \Pi^{\perp}_{M_{CM}} \bullet K + \Pi^{\parallel}_{M_{CM}} \bullet \Pi^{\perp}_{U_0} \bullet K + \Pi^{\parallel}_{M_{CM}} \bullet \Pi^{\parallel}_{U_0} \bullet \Delta . \tag{4.5}$$

Eq. (4.5) is graphically represented in Fig. 4.3. Both K and Δ are partitioned by the 4 projection operators into complementary parts. Depending on the content of U^0, the final residual error E consists of different components of K and Δ, as shown in Fig. 4.4.

- *Alignment biased*: This corresponds to the case where U^0 is empty, thus the input orbit is completely blamed on monitor errors. The interpreted orbit is identically zero and basically no orbit correction is necessary. In a more relaxed variation U^0 can contain only injection errors and steering on the interpreted orbit amounts to an injection fix based on the global fitting of the orbit to injection coordinates.
- *Monitor biased*: This corresponds to the case where U^0 spans the entire orbit space, thus the input orbit is completely blamed on alignment errors. The interpreted orbit is identical to the apparent orbit and orbit correction is done to eliminate as much as possible the apparent BPM pattern. This is what normally happens with simple-minded steering.

Alignment bias reflects an emphasis on the global consistency, favoring long range orbit pattern over local deviations, whereas the monitor bias prefers local flexibility, bending trajectory to fit monitor readings wherever needed. In the intermediate case U^0 takes on a finite subset of M^{EM}, allowing trade-off between the two contributions. For example, it is often found that a minimal subset of vectors in M^{EM} contributes the most toward reducing the residual T_Δ of Eq. (4.4). This minimal subset is then a natural choice for U^0.

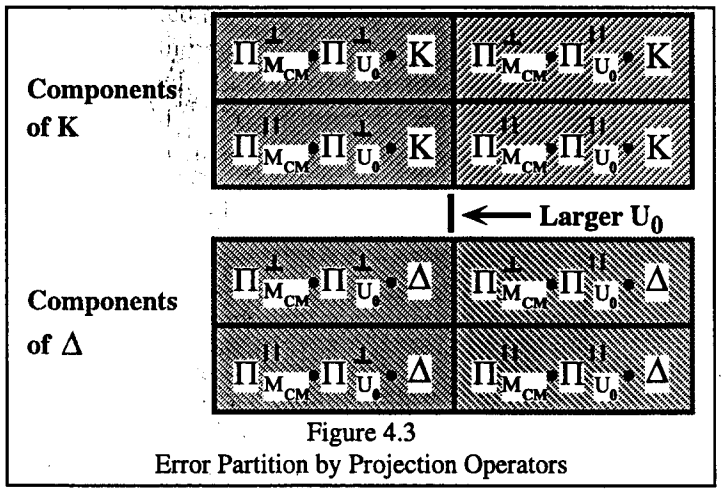

Figure 4.3
Error Partition by Projection Operators

[11] The term "interpretation" is already a hint of the highly under-constrained nature of the problem. In such cases intelligence of the algorithm used for interpretation is crucial.

4.1.2 *Including monitor errors as virtual correctors*

We can expand the scope of virtual correctors of Eq. (4.4) further to include monitor errors. The latter, instead of passively falling off as the residual orbit of steering by virtual correctors, can now actively match to the orbit pattern as a distinctive error signature. This is done by augmenting M^{EM} with columns making up the $N_M \times N_M$ identity matrix, N_M being the number of monitors. We denote this new response matrix M^{EMM}. The residual orbit after steering by M^{EMM} will be attributed to noise only. This scheme allows the monitor error to be identified exactly as what it is, and not misinterpreted as a misalignment error. The net effect is a more accurate separation of K and Δ, thus a smaller E.

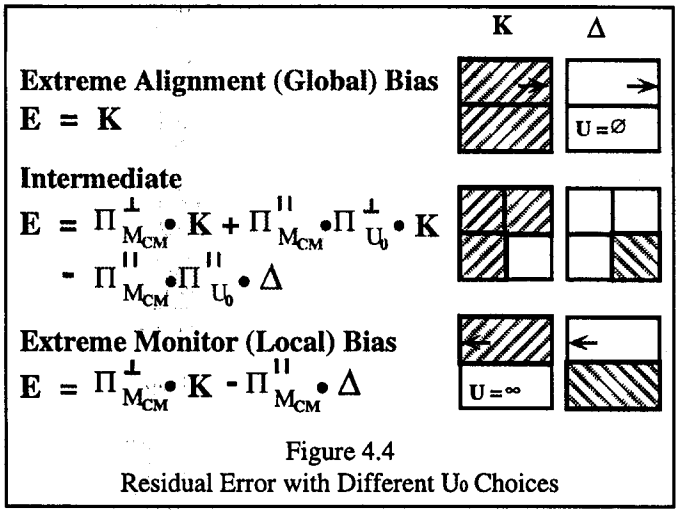

Figure 4.4
Residual Error with Different U_0 Choices

4.1.3 *Algorithm for optimizing U^0*

It is obvious by now that the element critical to the success of orbit interpretation is an intelligent algorithm which decides on the content of U^0 that disentangles K and Δ, such that the residual orbit E approaches the fundamental uncorrectability shown in Fig. 4.5.

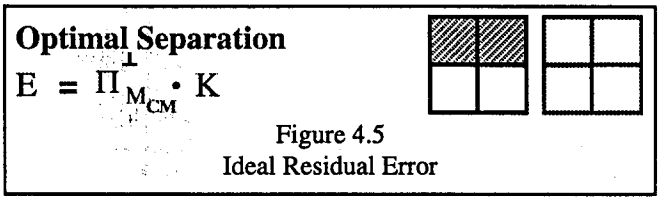

Figure 4.5
Ideal Residual Error

Due to the under-constrained nature of the problem, there is ample room for user preferences. For example, whether there is more confidence in alignment and field accuracy or monitor accuracy, whether emphasis is on global consistency or local flexibility, whether minimal total RMS or smallest number of errors is preferred in the makeup of U^0, etc. Relative weighting between different errors also reflects the user's bias, which ultimately depends on realistic evaluation of the steering situation. Once these preferences are set, however, it is up to an automated algorithm to decide on the content of U^0. A collection of algorithms are listed in Table 4.1, most of which have been tested for this purpose. Also listed are optimization biases of each algorithm. SVD and QR-decomposition are established mathematical algorithms. SUSMIC was locally developed for optimizing solutions for under-constrained systems. These three algorithms span the spectrum of preference between minimizing error RMS and error number. Being conceived as analytic and exact algorithms, they also share emphasis on minimizing response matrix singularity, and preclusion of noise cutoff as a physical option. These can be positive or negative features depending on the situation.

Table 4.1
Orbit Interpretation Algorithms

Algorithm	Alignment / Monitor Bias	Error RMS / Number Bias	Global / Local Bias
SVD	weighted	RMS	variable
QRD	weighted	mixed	variable
SUSMIC	none	number	variable
Prototype 3	variable	RMS	mixed
PROSAC	none	mixed	variable
MICADO	none	number	variable

Prototype 3 is an iterative algorithm developed exclusively for orbit interpretation and tested on real data at CEBAF [4]. It applies a prioritized sequence of SVD-generated alignment error combinations to the orbit. At each successive application the residual orbit is used to update a weighting factor on the BPM's such that distinct monitor errors can stand out above noise. The iteration terminates when an unnatural jump is detected in the total magnitude of alignment errors,

signaling monitor errors being misinterpreted as alignment-induced. The algorithm also provides continuous interpolation between alignment and monitor biases, and automatic optimization on this interpolation. It has been applied to real data at CEBAF and successfully demonstrated isolation of monitor errors from misalignment errors. Piecewise concatenation of alignment-biased and monitor-biased interpreted orbits has also been experimented as steering input, showing advantage in special cases, but automation appeared difficult.

Recently attention has been turned to MICADO as the orbit interpretation engine. Since MICADO was conceived as a corrector selection algorithm, it possesses two unique advantages. Firstly it aims at minimizing total number of errors instead of total error RMS, avoiding smearing out localized error effect and mixing monitor error into alignment error. Secondly it allows for residual noise after correction, avoiding exaggeration of errors that happens with exact algorithms. PROSAC, with its corrector prioritization feature, may also prove effective as an orbit interpretation engine in minimizing number of errors. There has been no test with MICADO or PROSAC at this point.

4.1.4 Orbit interpretation for multiple pass orbit

Orbit interpretation can be applied to multiple pass orbit sharing common alignment and monitor errors. Here the problem shifts away from being under-constrained and may actually become over-constrained, greatly improving the predictive power of orbit interpretation. This was indeed the case with the simultaneous multiple pass linac steering described in Section 3 [5].

4.2 Virtual monitors

As indicated in Fig. 4.2, the virtual monitor algorithm performs the following functions.

4.2.1 Generalized steering constraints

Once the interpreted orbit is established, we have knowledge of all beam coordinates at all locations, whether directly observable or not. Steering constrints on these coordinates can be easily output to a generic steering engine disguised as ordinary constraints at "virtual" monitors. In this scheme we can even constrain the absolute values of these coordinates if there is enough confidence in the interpreted orbit. The constraint on orbit-dependent dispersion is realized through

$$\delta M_{i6}^{ab} = 2 \sum_{j\neq 6} \sum_{k} \sum_{a\le c\le b} T_{ijk}^{cb} M_{j6}^{ac} \delta X_k^c + 2 \sum_{j} \sum_{a\le c\le b} T_{ij6}^{cb} \delta X_j^c \qquad (4.6)$$

where δM is the change in dispersion from location a to location b due to orbit changes δX at all locations c between a and b. M and T are first and second order optical transfer elements. Thus δM^{ab} can be directly input to the steering engine as a virtual orbit with associated response matrix being the right hand side sums of Eq. (4.6) with δX^c replaced with real response matrix elements M_{j2}.

4.2.2 Controlling response matrix singularity

After the complete set of response matrices is assembled, connecting correctors to both real and virtual monitor inputs, the algorithm enforces singularity control in the form of yet more virtual monitors. This is done by automated placement of virtual monitors at strategically chosen locations coupling strongly to singular combinations of correctors. This is a better way of controlling singularity than limiting or elimination of correctors, because it

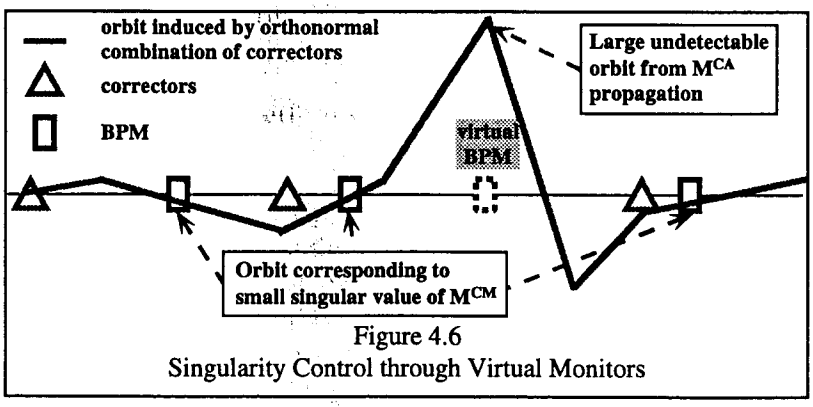

Figure 4.6
Singularity Control through Virtual Monitors

is always the singular combination of correctors, not individual ones, that causes steering problem. Indiscriminate limiting or elimination of individual correctors targets harmless and offending combinations equally, thus compromising steering effectiveness. The advantage of singularity constraint through virtual monitors is especially apparent in cases of disabled BPM's.

To do this, the program performs SVD on M^{CM} for its condition number, and calculates its Gram determinant. Both are compared to numerical criteria. If either criterion is not met, the SVD generated corrector combination with the smallest singular value is identified. The corrector-to-all-locations matrix M^{CA} is applied to this combination. The index of the largest component of the out-come vector points to the location for a new virtual monitor, with its orbit given by the interpreted orbit. Iteration stops when the system passes both numerical tests. The procedure can be visualized from Fig. 4.6.

At CEBAF steering with virtual monitors has been experimented in the spreaders, extraction region, and arcs. It demonstrated ability to bypass monitor errors in orbit correction, and good singularity control without unnecessary compromise of corrector strengths. It has also been tested for angle control in the spreaders. The entire package of orbit interpretation and virtual monitor algorithms has not been developed for routine operation at this point.

4.3 Application to orbit reproduction

The majority of steering effort in accelerator operation goes into reproducing an established standard orbit, where the goal is to minimize the difference between the standard and absolute orbits, rather than the absolute orbit itself. In such operations the orbit interpretation and virtual monitor algorithms are even more effective since the relative orbit usually exhibits effects due to far fewer errors than the absolute orbit. Clean separation between error sources by orbit interpretation is much more achievable, and the outcome can be very useful for machine diagnosis. Extension to relative orbit at virtual monitors is straightforward. This guarantees a much more complete orbit reproduction than is possible with only real monitors.

Acknowledgements

The author would like to acknowledge very helpful inputs from Andrew Hutton on many topics discussed in this report, and substantial contribution by Johannes van Zeijts toward both realizing and improving the orbit correction algorithm at CEBAF. The author would also like to acknowledge either help or input from the following people: Walt Akers, Joseph Bisognano, Alex Bogacz, Bruce Bowling, David Bryan, David Douglas, Leigh Harwood, Valeri Lebedev, Hamid Shoaee, Chip Watson, and Sue Witherspoon.

References

[1] B. Autin and Y.Marti, *Closed orbit correction of alternating gradient machines using a small number of magnets*, CERN ISR-MA/73-17 (1973).

[2] W. Press, B. Flannery, S. Teukolsky & W. Vetterling, *Numerical Recipes in C*, Cambridge University Press, 1988.

[3] Y. Chao, *Methods of Orbit Correction System Optimization*, Proceedings of 1997 Particle Accelerator Conference, Vancouver, Canada.

[4] Y. Chao et al, *Orbit Correction Using Virtual Monitors at Jefferson Lab*, Proceedings of 1997 Particle Accelerator Conference, Vancouver, Canada.

[5] Y. Chao et al, *Simultaneous multiple pass steering at Jefferson Lab*, to be submitted to the 1999 Particle Accelerator Conference, New York, USA.