Large Hadron Collider Project

LHC Project Report 470

Analysis and Optimisation of Orbit Correction Configurations Using Generalised Response Matrices and its Application to the LHC Injection Transfer Lines TI 2 and TI 8

Y. - C. Chao[#], V. Mertens

Abstract

The LHC injection transfer lines TI 2 and TI 8 will transport intense high-energy beams over considerable distances. In their regular part a FODO lattice is used with 4 bending magnets per half-cell and a half-cell length of 30.3 m, similar to that of the SPS. The relatively tight apertures in these lines require precise trajectory control. Following an earlier study a baseline correction scheme was chosen where two out of every four consecutive quadrupoles are complemented with correctors and beam position monitors ("2-in-4"). With the ordering of the equipment approaching, a further in-depth investigation has been made using a newly developed analytic method. This method evaluates, based on the design specifications, the global performance of an orbit correction system in terms of observability, correctability, correction range and response singularity. In addition, orbit and error envelopes are obtained over the full beam line in an efficient and rigorous manner, providing insights not easily accessible with conventional tools. The cost/performance ratio of a given configuration can be optimised, both analytically through the elimination of structural defects and numerically through fine-tuning. Finally, features for failure mode analysis allow the user to diagnose observed performance anomalies, and features for critical-element analysis enable the user to identify weak spots in the configuration. The method is described in detail to facilitate the interpretation of the results obtained for TI 2 and TI 8, and to allow their application to other orbit correction systems. The new, optimised 2-in-4 scheme permits some hardware economies at comparable performance. Further exploration has identified an alternative scheme with a 1-in-3 corrector and 2-in-3 position-monitor pattern. At an overall cost comparable to the 2-in-4 scheme this latter configuration maintains the possibility of intuitive one-to-one correction, important in the commissioning phase, at a performance slightly above the nominal aperture budget, but allows to reduce, using computer support, the corrected maximum trajectory excursions significantly below those of the 2-in-4 scheme.

SL Division

[#] On leave from Thomas Jefferson National Accelerator Facility, 12000 Jefferson Avenue, MS-5A, Newport News, VA 23606, USA. This work was supported in part by the U.S. DOE Contract No DE-AC05-84-ER40150.

Administrative Secretariat
LHC Division
CERN
CH - 1211 Geneva 23
Switzerland
Geneva, 17 May 2001

Table of Contents

1 Introduction

1.1 The LHC Injection Transfer Lines TI 2 and TI 8

The new LHC injection transfer lines TI 2 and TI 8 have a combined length of about 5.6 km and use over 700 room-temperature magnets. The geometrical location of TI 2 and TI 8 is given in Figure 1.

In the regular part of both lines a FODO lattice with about 90° phase advance per cell is used with 4 bending magnets per half-cell and a half-cell length of 30.3 meters, similar to that of the SPS. Figure 2 shows the sequence of elements with betatron and dispersion functions for both lines. The main optical parameters and requirements are summarised in Table 1. The main horizontal arc in TI 2 has been designed as an achromat. Space reasons dictated a different solution for TI 8. Beam optics calculations to second order show negligible effects, which do not require higher-order corrections. A more detailed description of these lines is given in [1].

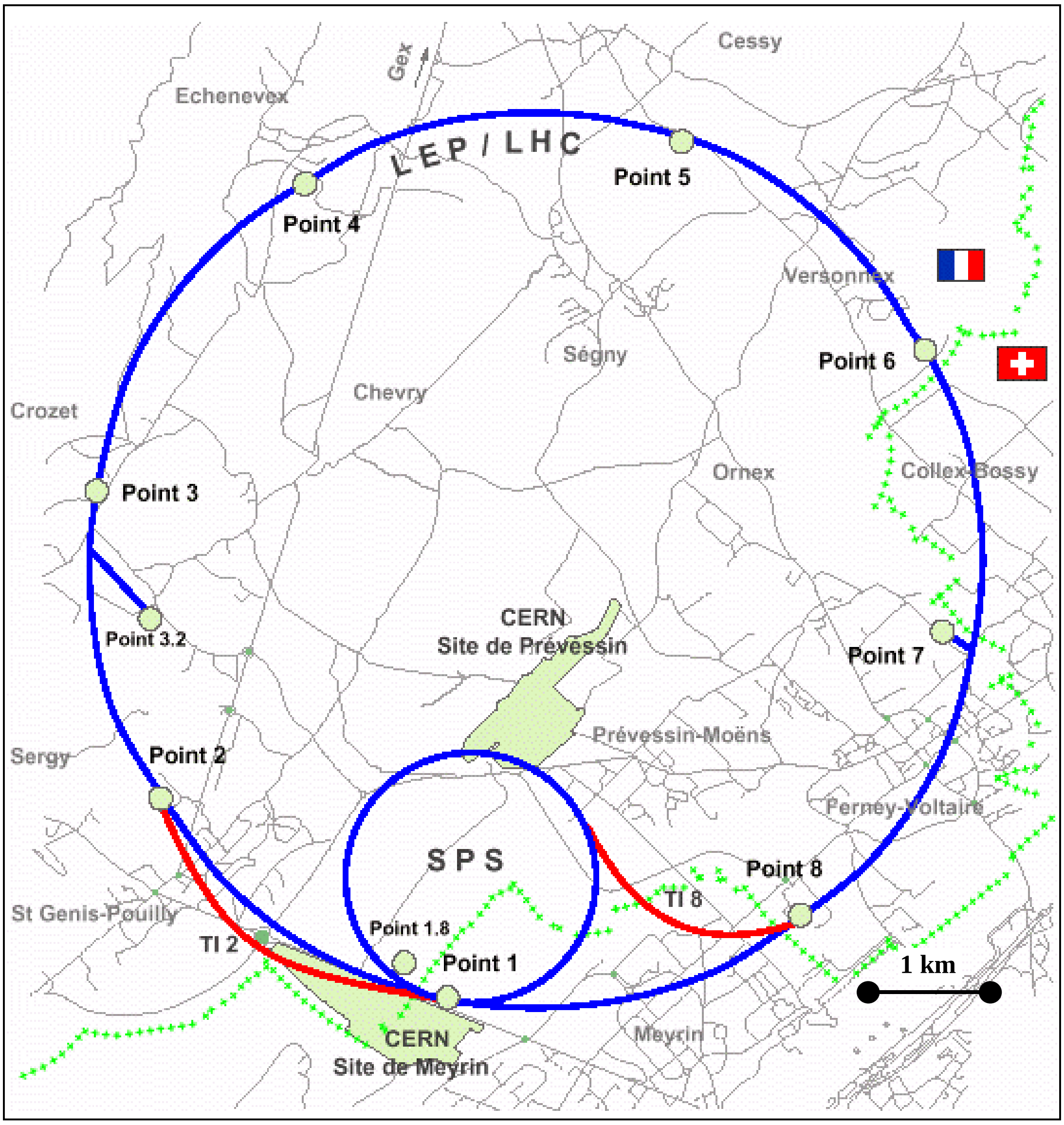


Figure 1: Location of TI 2 and TI 8 with respect to SPS and LHC.

The proton beam parameters (energy 450 GeV; nominal intensity 3.17*10^{13} p per SPS cycle; nominal transverse emittance 3.5 µm*rad (RMS, normalised)) dictate that the beams must stay within the available aperture to avoid severe damage. The strongest aperture constraint comes from the new main dipoles with their full gap height of 25 mm which leads to a maximally tolerable vertical trajectory excursion, near the defocusing quadrupoles, of ±4.5 mm.

1.2 Earlier Study of Correction Schemes

A first study of possible correction schemes has been carried out in 1997 [2] which led to the adoption of a so called “2-in-4” configuration as the baseline scheme. In this scheme two out of every four consecutive quadrupoles, separately for each plane, are complemented with correctors. For maximum sensitivity the monitors giving the position feedback are placed at 90° phase advance, one cell downstream of the correctors. In the matching sections a full correction scheme was implemented to improve handling of extraction errors and to permit precision delivery into the LHC. The resulting configuration employed, for both lines together, a total of 110 corrective elements, of which 94 were assumed at that time to consist of modified LEP correctors.

The data presented in [2] had been obtained using a computer code initially written for particle tracking which had been extended to allow local trajectory correction in transfer lines [4], i.e. steering with one corrector and observing the effect with one monitor (“1-to-1 steering”). The exploration of the possible error combinations had been done through a large number of runs (typically 1000), using each time a different random pattern of errors within predefined error limits. The maximally occurring trajectory excursion after correction was then derived by taking the extrema over all runs. By its nature this method is quite time-consuming and not very intuitive in finding the truly optimum configuration in reasonable time.

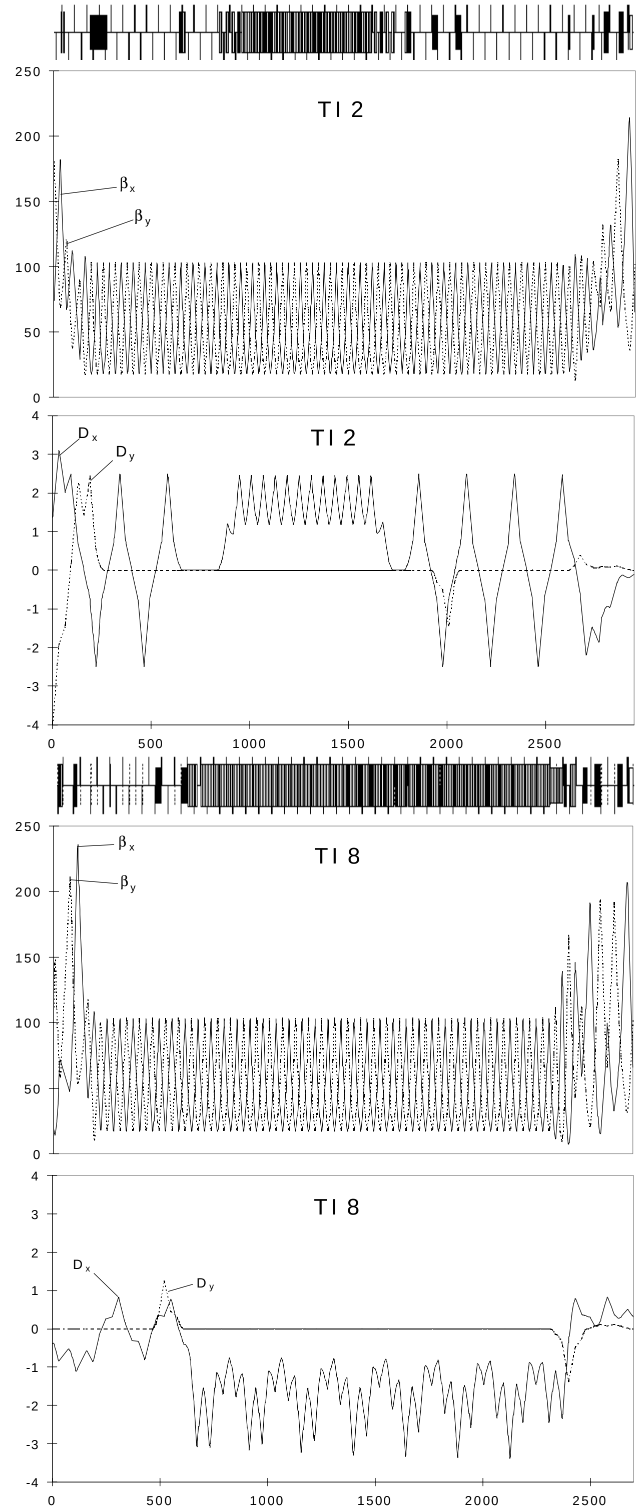


Figure 2: Sequence of elements, betatron and dispersion functions (all values in [m]) for TI 2 and TI 8.

1.3 Present Study

Item	TI 2	TI 8	Unit
Optics			
$\beta_{x,max}$	214.2	235.6	m
$\beta_{y,max}$	182.2	211.2	m
$\beta_{x,max,lattice}$	102.9	103.1	m
$\beta_{y,max,lattice}$	102.9	103.6	m
$D_{x,max}$	3.10	3.38	m
$D_{y,max}$	3.98	1.38	m
$\mu_{x,total}$	11.3	10.8	2π
$\mu_{y,total}$	11.4	10.4	2π
Half cell length	30.3	30.3	m
Number of half cells	95	85	
Acceptance			
Norm'd nominal emittance (β^2/σ)	3.5	3.5	µm
Assumed beam size	±4	±4	σ
Nominal momentum spread	±0.12	±0.12	%
Injection precision			
Deposition precision on LHC c.o.	±1.5	±1.5	σ
including:			
SPS c.o. errors at extraction			
Power supply ripple / drifts			
Injection kicker ripple / drifts			

Table 1: Main optical parameters and requirements of TI 2 and TI 8.

With the moment of freezing the layouts and ordering the equipment approaching it was decided to take another look into TI 2 and TI 8 trajectory correction, taking new aspects and modifications since the previous study into account. One major impact comes from the decision not to re-use the LEP correctors for this purpose. It turned out that, besides having to reduce their enormous gap by pole pieces, their coils would also need to be replaced as a result of the high irradiation in the last years of LEP operation, which dwarfs the economic advantage of a re-use. Instead, a new type of corrector is being conceived [3] offering a significantly greater bending power of 0.12 [Tm].

Despite that their impact on the overall results is pretty small, this new study also offered the opportunity to incorporate all design changes in TI 2 and TI 8 which had occurred meanwhile (choice and positioning of certain magnetic elements, new short straight section layout, optics fitted to LHC V6).

In contrast to the previous investigation the present study was performed using a much more systematic and graphically supported, thus more intuitive and efficient method. The algorithms have their origin in a previous program [5] consisting of recipes for incremental configuration improvements with a focus on observability, correctability, correction range, response singularity and economy of elements. These algorithms have been substantially expanded for the current analysis to establish correspondence to the performance criteria adopted in the previous study. The emphasis of the latter on obtaining distributions of final orbit errors for a given distribution of built-in and dynamic errors is echoed in the expanded analytic program, which quantifies various error and orbit distributions at all locations of the beam line through analytical methods. Besides retaining all previous features, the expanded program also added features related to failure-mode and critical element analyses.

To the extent that direct comparisons can be made with the previous analysis, very good agreement is demonstrated.

2 Methods of Analysis and Optimisation

The analytic methods used to arrive at optimised corrector-monitor configurations are discussed in the following in a rather condensed fashion, without too much digression into explicit mathematical details. It is safe to say that the collection of a few key generalised response matrices characterising the machine optics, the errors, and the orbit correction configuration, contains all there is to know about the performance of an orbit correction system. Furthermore, analytical methods in linear algebra, multi-dimensional vector analysis, optimisation theory, and probability distribution theory can be brought to bear on this collection of matrices, yielding quantitative assessments on its performance to an impressive degree of completeness, leaving only few esoteric questions unanswerable with mathematical tools available today. The construction of the generalised response matrices and the application of the analytical methods will be described in the following sections.

2.1 Generalised Response Matrices

Let us first define the elements that a response matrix connects to. These elements are usually termed “actuators” and “responders”. An actuator A imparts an action (e.g., a magnetic kick, a coordinate shift) which has an effect on a responder R (e.g., a position change, an angle change) described by a response coefficient C^{RA}, as

$$R = C^{RA} A. \tag{2.1}$$

When there are more than one actuator or responder, (2.1) can be generalised to the matrix form

$$\begin{pmatrix} R^1 \\ R^2 \\ \vdots \\ R^n \end{pmatrix} = \begin{pmatrix} C^{11} & C^{12} & \cdots & C^{1m} \\ C^{21} & C^{22} & \ddots & C^{2m} \\ \vdots & \ddots & \ddots & \vdots \\ C^{n1} & C^{n2} & \cdots & C^{nm} \end{pmatrix} \cdot \begin{pmatrix} A^1 \\ A^2 \\ \vdots \\ A^m \end{pmatrix} \tag{2.2}$$

with n responders and m actuators where the response coefficients C^{ij} are indexed by the corresponding actuator and responder in (2.2). As defined, the C^{ij} can thus take on values of elements in the ordinary optical transfer matrices (e.g., M^{16}, the dispersion, if the actuator is an energy offset and the responder is a position) or a combination thereof (e.g., sum over several M^{22}‘s, if the actuator is a “group of bending magnets” powered in series and the responder is an angle).

When the actuators are the corrector kick angles and responders the positions at monitors, the response matrix is just the typical one used in automated orbit correction. Our ability to evaluate the performance of an orbit correction system lies in allowing other types of entities to play the roles of either actuator or responder to form generalised response matrices.

Three groups of generalised actuators have been identified, listed in Table 1. The **alignment-error** actuators consist of deviations from the baseline design in terms of injection coordinate offset, alignment errors, or undesired magnetic field kicks. They can arise from the errors associated with a long dipole and thus acquire “position offsets” at the end of these dipoles, in addition to injection “position” errors. The **all-element** actuators consist of a representative subset of all the elements in the beam line, serving as a complete representation of the response property of the beam line, and a repository for candidate elements to be added to the configuration. The **corrector** actuators consist of the familiar correction elements, with the possibility of acquiring “position” effects for long dipoles, and can exert combined effects in the case of correction elements powered in series.

Likewise, we have identified two groups of generalised responders, again listed in Table 1. The **all-element** responders play the important role of monitoring the response of the entire beam line to the impact of all types of errors and orbit corrections. They are made up of a representative subset of all the elements in the beam line, possibly enhanced with angle coordinates for complete coverage of the phase space. The **monitor** responders are the familiar set of positions at monitors, possibly enhanced with angle coordinates. The responder coordinates can, in principle, be extended to a much wider set of entities [5], but to limit the focus to that relevant to the current report, we will not discuss this here. The discussion of the last entry in Table 1, $\mathbf{M^{MM}}$, is deferred until the next section.

Generalised resp. matrix	**(Generalised) actuator**	**(Generalised) responder**	**Response coefficients**
$\mathbf{M^{CM}}$	$\mathbf{A^C}$: correctors, dipoles, dipole strings	$\mathbf{R^M}$: position & angle at monitors	M^{11},M^{12},M^{21},M^{22} and linear comb.
$\mathbf{M^{EM}}$	$\mathbf{A^E}$: alignment type errors (injection, misalignment, field, ….)	$\mathbf{R^M}$: position & angle at monitors	M^{11},M^{12},M^{21},M^{22}
$\mathbf{M^{CA}}$	$\mathbf{A^C}$: correctors, dipoles, dipole strings	$\mathbf{R^A}$: position & angle at all representative elements	M^{11},M^{12},M^{21},M^{22} and linear comb.
$\mathbf{M^{EA}}$	$\mathbf{A^E}$: alignment type errors (injection, misalignment, field, …)	$\mathbf{R^A}$: position & angle at all representative elements	M^{11},M^{12},M^{21},M^{22}
$\mathbf{M^{AM}}$	$\mathbf{A^A}$: angle at all representative elements	$\mathbf{R^M}$: position & angle at monitors	M^{12},M^{22}
$\mathbf{M^{MM}}$	$\mathbf{A^M}$: monitor offset error	$\mathbf{R^M}$: apparent orbit error at monitor	δij (see Section 2.3)

Table 1: Generalised response matrices

Table 1 gives the generalised actuators and responders associated with each generalised response matrix in the sense of (2.2) Thus, for example, for the matrix $\mathbf{M^{EA}}$, we can construct the vector representing all alignment type errors $\mathbf{A^E}$, which, when acted on by $\mathbf{M^{EA}}$, yields the vector $\mathbf{R^A}$ representing the position (and/or angle) errors at all representative locations in the line. Similarly, one can establish the vectors $\mathbf{A^C}$, $\mathbf{A^A}$, $\mathbf{A^M}$, and $\mathbf{R^M}$**,** respectively, to represent the impact from corrector (including dipole) kicks, all (candidate) elements, monitor errors, and response at monitors. The relations between these quantities through the matrices in Table 1 and secondary matrices, to be discussed in later sections, constitute the core of the current analysis. The more basic relations can be readily made explicit below.

2.1.1 Alignment Type Error Effects

The impact of alignment type errors (injection, misalignment and unaccounted field errors), including angle effects, is manifested at all locations in the beam line, and in particular at the monitors through

$$\begin{aligned} \mathbf{R^A} &= \mathbf{M^{EA}}\cdot\mathbf{A^E} \\ \mathbf{R^M} &= \mathbf{M^{EM}}\cdot\mathbf{A^E} \end{aligned} \qquad (2.3)$$

2.1.2 Corrector Effects

The impact of corrector kicks (including “position” offsets caused by long dipoles and possible combined effects) on position (and/or angle) at all locations, and monitors in particular, can be expressed as

$$\begin{aligned}\mathbf{R}^{\mathbf{A}} &= \mathbf{M}^{\mathbf{CA}} \cdot \mathbf{A}^{\mathbf{C}} \\ \mathbf{R}^{\mathbf{M}} &= \mathbf{M}^{\mathbf{CM}} \cdot \mathbf{A}^{\mathbf{C}}\end{aligned} \tag{2.4}$$

2.1.3 Orbit Correction

The most straightforward realisation of an orbit correction process is through the pseudo-inverse of the matrix $\mathbf{M}^{\mathbf{CM}}$, denoted $\mathbf{M}^{\dagger}{}_{\mathbf{CM}}$:

$$\begin{aligned}\mathbf{A}^{\mathbf{C}} &= -\mathbf{M}^{\dagger}_{\mathbf{CM}} \cdot \mathbf{R}^{\mathbf{M}} \\ \mathbf{M}^{\dagger}_{\mathbf{CM}} &= \left(\mathbf{M}^{\mathrm{T}}_{\mathbf{CM}} \cdot \mathbf{M}^{\mathbf{CM}}\right)^{-1} \cdot \mathbf{M}^{\mathrm{T}}_{\mathbf{CM}}\end{aligned} \tag{2.5}$$

Equation (2.5) effectively implies a singular value decomposition (SVD) orbit correction scheme. However, as it can be seen from subsequent sections, the adoption of the pseudo-inverse matrices actually paves the way for discussions of observability, correctability, and underlying errors based on division of error and corrector vector spaces by projection operators (expressible as pseudo-inverses). This conclusion is actually inescapable regardless of the orbit correction scheme used. This is because the SVD method, when ignoring corrector limits, is guaranteed to maximally exploit the subspace defined by the projection operator, and thus sets an upper bound to maximally attainable orbit correction.

2.2 Mathematical Tools

Before embarking on the core analysis, it is beneficial to review a few important mathematical concepts heavily relied on in this analysis. The focus shall be on immediate physical contexts of these concepts, rather than to elaborate on details, which, although critical to the success of the analysis, have to be deferred until the Appendices. The same theme, applied to specific matrices below, will be played on other response matrices repeatedly in the later analysis to extract information about the orbit correction system.

It is necessary to justify here the extent to which we carried this analysis in applying mathematical tools to realise all the analysis in terms of linear problems. In principle modern numerical tools such as *Mathematica* can solve complicated non-linear systems of high order and dimensionality, such as formally presented by the current problems, before attempting simplification. However, for a problem of the scale as under study here, where the number of potential error sources and monitored points runs to the hundreds or even a thousand, repeatedly and blindly applying non-linear solvers or optimisers with dubious initial guesses is not an option, both in terms of reliability and efficiency. Instead, all the analysis must be realised in linear form and in terms of a limited set of realistic operations (see Appendix A), where efficient and robust algorithms can be applied, without compromising mathematical rigour. The emphasis here is therefore on a detailed description of recipes leading from intuitive pictures to the input end of well-defined numerical algorithms. These recipes also provide insight not possible through blind application of non-linear solvers. As will be seen, linear methods can carry this analysis indeed a long way.

Most algorithms developed for this report, although not contributing in any sense to the fundamental knowledge in linear algebra, vector calculus, and theories of optimisation or probability distribution, are nonetheless not readily found in standard textbooks or references under these subjects. This is probably due to their highly specialised focus on parochial problems such as encountered here. This again justifies a detailed treatment. They can be valuable to the analysis of other problems of similar nature. For the treatment of algorithms more readily found in general literature, such as SVD, we limit the discussion to the minimum.

2.2.1 Projection Operators, Null Space and Orthogonalising Transform

The projection operators divide a vector space into complementary subspaces spanned by collections of vectors. Taking the response matrix $\mathbf{M^{CM}}$, and construct the matrices

$$\begin{aligned} \Pi^{\parallel}_{\mathbf{CM}} &= \mathbf{M^{CM}} \cdot \mathbf{M}^{\dagger}_{\mathbf{CM}} = \mathbf{M^{CM}} \cdot \left(\mathbf{M}^{\mathsf{T}}_{\mathbf{CM}} \cdot \mathbf{M^{CM}}\right)^{-1} \cdot \mathbf{M}^{\mathsf{T}}_{\mathbf{CM}} \\ \Pi^{\perp}_{\mathbf{CM}} &= \mathrm{I} - \Pi^{\parallel}_{\mathbf{CM}}, \end{aligned} \tag{2.6}$$

where **I** is the identity matrix, then $\Pi^{\parallel}_{\mathbf{CM}}$ and $\Pi^{\perp}_{\mathbf{CM}}$, respectively, project the monitor space into the subspace spanned by the column vectors of $\mathbf{M^{CM}}$, i.e., those corresponding to “corrector effects”, and the subspace orthogonal to the first subspace. One sees immediately that these two operators divide the monitor space into the part correctable by correctors and that un-correctable.

It should be noted that (2.5) and (2.6) assume that the number of monitors is equal to or greater than that of the correctors, i.e., the system is critically or over-constrained. The application of projection operators in the under-constrained case has to be carried out differently and has a different significance. There is also extra complication when $\mathbf{M^{CM}}$ is degenerate. These cases will be discussed in Appendix B.

The null space of the matrix $\mathbf{M^{EM}}$ is the subspace in the error vector space spanned by all error vectors $\mathbf{A^{E}}$ satisfying

$$\mathbf{M^{EM}} \cdot \mathbf{A^{E}} = 0. \tag{2.7}$$

One can also construct a “null space matrix” $\mathbf{M^{EM}_{null}}$ whose rows form an orthonormal basis of all such $\mathbf{A^{E}}$s, called “null space vectors”. It is clear, for example from (2.7), that $\mathbf{M^{EM}_{null}}$ contains all error combinations that are not observable at the monitors.

One can further find an orthonormal transformation from the original error space[1] to a new basis where the null space vectors are part of the basis vectors, with the remaining basis vectors made of those not satisfying (2.7). An efficient way to do this is given in Appendix C. The advantage of such a transform is that it neatly breaks up response matrices such as $\mathbf{M^{EM}}$ into observable and unobservable parts, while keeping the shape of the error distribution intact if it was already normalised.

2.2.2 Multi-Dimensional Ellipsoid, its Projection onto Higher and Lower Dimensions, and the Inverse Projection of Solution(s)

The error distribution for the current analysis must satisfy the following criteria:

- All errors have, or can be so grouped as to have, <u>independent</u> probability distributions.
- The overall probability <u>density</u> of any combination of such errors, represented by the vector $\mathbf{A^{E}}$, is a function $\mathbf{P_E}$ of a quadratic form **Q** in the error vectors only.
- The quadratic form **Q**, when represented as a square matrix, is <u>symmetric</u> and has only <u>non-negative</u> eigenvalues.

[1] preferably normalised

- The function $\mathbf{P_E}$ is normalisable (integrates to a finite number), although not necessarily monotonically decreasing.

This analysis will be valid as long as the above conditions are met. In practice one has here to deal mainly with independent Gaussian distributions, one for each error source with distinct σ's, with overall probability density being the product of constituent ones. Obviously, this satisfies the above criteria. On the other hand, there is no need to explicitly invoke the Gaussian distribution for the discussion of this section.

Given an error pattern represented by a vector $\mathbf{A^E}$, the above criteria is taken to imply that

$$\begin{aligned} &\mathbf{P_E} = \mathbf{P_E}(\mathrm{Q}) \\ &\mathrm{Q} = \mathbf{A}^{\mathrm{E}^{\mathrm{T}}} \cdot \mathbf{E}^{\mathrm{E}} \cdot \mathbf{A}^{\mathrm{E}}. \end{aligned} \tag{2.8}$$

The square matrix $\mathbf{E^E}$ then must be symmetric and have only non-negative eigenvalues. Physically (2.8) represents a probability distribution with contours of constant probability defined by concentric convex ellipsoidal surfaces, the principal axes of which are determined by eigenvectors of $\mathbf{E^E}$. The symmetry and convexity of $\mathbf{E^E}$ are what make the vast majority of the subsequent analysis possible.

The contour of constant probability density, and thus the total accumulated probability, represented by the multi-dimensional volume enclosed in this contour, can undergo a multitude of transformations via the primary response matrices, or the secondary response matrices to be discussed later. During such transformations the total **accumulated** probability contained in the contours in different vector spaces connected by these transformations does not change. This means that, if we can identify rules for mapping these ellipsoidal contours between different vector spaces, we can start with the orthonormalised probability distribution in the error space, and, through various transformations, examine the probability distributions in other primary or secondary spaces.

We will now briefly describe the concept of mapping an ellipsoid from space **A** to space **B** via a matrix **M**, and formal equations describing these concepts. Complete mathematical details are given in Appendix D.

- Mapping between spaces with equal dimensionality and no rank deficiency in **M**: In this case mapping of $\mathbf{E^E}$ is as straightforward as matrix multiplication. All that is needed is the inverse of **M**, which by definition exists.

$$\mathbf{E}^{\mathrm{E}} \rightarrow \mathbf{M}^{\mathbf{-1}^{\mathrm{T}}} \cdot \mathbf{E}^{\mathrm{E}} \cdot \mathbf{M}^{\mathbf{-1}}. \tag{2.9}$$

- Mapping into lower dimension (including rank deficiency in **M**): The final mapped ellipsoid in **B** is determined by a sub-contour on **A** which is locally perpendicular to the null-space vectors of **M**. The mapped ellipsoid is an intersection between **B** and a "hyper-cylinder" of null-space vectors containing this sub-contour of **A**.

$$\begin{aligned} &\mathbf{E''}^{\mathrm{E}} = \mathbf{T}^{\mathrm{T}} \cdot \mathbf{E'}^{\mathrm{E}} \cdot \mathbf{T}, \\ &\mathbf{T}^{\mathrm{K}} = \mathbf{I} \oplus \mathbf{K}, \\ &\mathbf{K} = -\left(\mathbf{E}^{\mathrm{E}}_{\mathrm{P}}\right)^{-1} \cdot \mathbf{E}^{\mathrm{E}}_{\mathrm{R}}, \end{aligned} \tag{2.10}$$

where $\mathbf{E_P}$ and $\mathbf{E_R}$ are matrices that together make up the envelope condition in an orthonormalised space of **A** based on (D.3).

- Mapping into higher dimensions: The image space of **M** in **B**, defined by the complement to the null space of **M**, must be identified first. The map $\mathbf{M^P}$ from **A** to the image space can then proceed as in the first case, with additional constraints defined by the equations governing the image space.

$$\mathbf{E^E} \rightarrow \mathbf{M^{p\text{-}1^T}} \cdot \mathbf{E^E} \cdot \mathbf{M^{p\text{-}1}}. \tag{2.11}$$

Equally important is the need to find the point(s) **Z** on the original ellipsoidal contour in **A** that is (are) mapped into specific point(s) **X** in **B** by **M**. This is important since after identifying an extremum in the mapped space one often needs to know, for example, what error configuration(s) caused it. This inverse process is also quite involved and differs between the modes of mapping mentioned above. We will describe the formal concept here, using a minimum of equations, and defer again the details to Appendix E.

- Back mapping between spaces with equal dimensionality and no rank deficiency in **M**: This is again a trivial case yielding

$$\mathbf{Z} = \mathbf{M^{-1}} \cdot \mathbf{X}. \tag{2.12}$$

- Back mapping into higher dimension (including rank deficiency in **M**): The inversely mapped point is determined by two sets of equations

$$\begin{aligned}
&\sum_{k} \mathrm{M}^{ik} \cdot \left(\mathrm{Z}^{k} - \mathrm{Y}^{k}\right) = 0, \quad \mathrm{i} = 1,2,\ldots.\mathrm{Nr.} \\
&\nabla_{\mathrm{N}}^{\mathrm{i}} \mathrm{S}\big|_{\mathrm{Z}} = 2\,\mathrm{N}^{\mathrm{i}} \cdot \mathrm{E} \cdot \mathrm{Z} = 0, \qquad \mathrm{i} = \mathrm{Nr}+1,\ \mathrm{Nr}+2,\ldots\mathrm{Nc.} \\
&\mathrm{Y} = \mathrm{M}^{\dagger} \cdot \mathrm{X}\ .
\end{aligned} \tag{2.13}$$

where Nr and Nc are the row and column dimensions of **M**, and **N** consists of the null space vectors of **M**.

- Back mapping into lower dimension: This is uniquely determined by

$$\mathbf{Z} = \mathbf{M}^{\dagger} \cdot \mathbf{X}. \tag{2.14}$$

2.2.3 Gradient, Inscribing Points, Extrema of Arbitrary Operators on a Constrained Contour, Hessian, and Local Curvature

For the probability distribution $\mathbf{P_E}$ described in the last section, with the equation for the contour of constant probability density given by (2.8), one can easily derive the gradient vector at a given point E normal to the contour

$$\begin{aligned}
\nabla \mathrm{Q}\big|_{\mathrm{E}} &= \sum_{\mathbf{ijk}} \hat{\mathbf{x}}_{\mathbf{i}} \left.\frac{\partial A_{\mathbf{j}}^{\mathbf{E}} A_{\mathbf{k}}^{\mathbf{E}} E_{\mathbf{jk}}^{\mathbf{E}}}{\partial A_{\mathbf{i}}^{\mathbf{E}}}\right|_{\mathrm{E}} = 2 \sum_{\mathbf{ijk}} \hat{\mathbf{x}}_{\mathbf{i}} \delta_{\mathbf{ij}} A_{\mathbf{k}}^{\mathbf{E}} E_{\mathbf{jk}}^{\mathbf{E}}\Big|_{\mathrm{E}} \\
&= 2 \sum_{\mathbf{ik}} E_{\mathbf{ik}}^{\mathbf{E}} A_{\mathbf{k}}^{\mathbf{E}}\, \hat{\mathbf{x}}_{\mathbf{i}}\Big|_{\mathrm{E}} = 2 \cdot \mathbf{E^E} \cdot \mathbf{A^E}.
\end{aligned} \tag{2.15}$$

In (2.15), we have used $\hat{\mathbf{x}}_\mathbf{i}$ for the unit vector in the i-th direction and *italic* letters for the components of $\mathbf{A}^\mathbf{E}$ and $\mathbf{E}^\mathbf{E}$. Likewise, we have used the symmetry of $\mathbf{E}^\mathbf{E}$ to arrive at (2.15). The final expression is to be understood as a vector in the error space.

The gradient vector of (2.15) is critical, together with the Lagrange multiplier method, in obtaining local tangency to hyper-planes and extreme values on the ellipsoid surface. Again, we describe the formal concepts and formulae below, deferring further details until the Appendices F and G.

- Inscribing a point of the ellipsoid **E** to the hyper-plane equations $\mathbf{P}^\mathbf{i}(\mathbf{X})=\pm\mathbf{V}_\mathbf{i}$: This is given by coincidence of the gradient vectors of **E** and **P**

$$\begin{aligned}&\sum_\mathbf{k}\mathbf{M}_\mathbf{i}^\mathbf{k}\cdot\mathbf{X}^\mathbf{k}=\mathbf{P}^\mathbf{i}(\mathbf{X})=\pm\mathbf{V}_\mathbf{i},\\&2\mathbf{E}\cdot\mathbf{X}=\lambda_\mathbf{i}\mathbf{M}_\mathbf{i}.\end{aligned}\tag{2.16}$$

- Extreme values of an operator **Π** on the surface constrained by an ellipsoid equation **E**=S: This can be solved by the eigenvalue equation

$$\left(\mathbf{\Pi}^\mathrm{T}\cdot\mathbf{\Pi}\cdot\mathbf{E}^{-1}-\lambda\right)\cdot\mathbf{E}\cdot\mathbf{X}=\mathbf{0}.\tag{2.17}$$

- Hessian and local curvature: Once an interesting point is identified on the ellipsoid (e.g., a point of tangency to a plane, or an extremum), we can use the concept extended from the linear gradient to derive curvature estimates, which indicate the persistence of the quality (extremum, tangency, etc.) of interest. These extensions involve the Hessian $\mathbf{H}_\mathbf{E}$ of the ellipsoid, and the second derivative of **E** taken along the direction of the gradient at the point of interest. The "rate of recession" of the ellipsoid around such points is given by

$$\begin{aligned}&\mathrm{R}=\frac{|\mathrm{V}|^2\mathrm{Det}\left(\mathbf{E}^\mathrm{T}\cdot\mathbf{E}\right)}{\mathbf{X}^\mathbf{T}\cdot\mathbf{E}^3\cdot\mathbf{X}},\\&|\mathrm{V}|^2=\mathbf{X}^\mathbf{T}\cdot\mathbf{E}^2\cdot\mathbf{X}.\end{aligned}\tag{2.18}$$

2.2.4 Properties Specific to The Multiple Gaussian Distribution

We can advance the analysis further if we restrict the generic probability distribution used in the previous sections to define the ellipsoidal constant-probability surface further to be Gaussian in all independent dimensions. Firstly it is known that such a multiple dimensional Gaussian distribution remains a multiple dimensional Gaussian under a linear transformation. This allows direct translating of vectors from one vector space to another with correlated probability density.

The N-dimensional probability density resulting from N <u>independent</u> 1-dimensional Gaussian distributions is simply

$$\begin{aligned}&P(x_1,x_2,\ldots x_\mathrm{N})=\prod_\mathrm{i}^\mathrm{N}\frac{1}{\sigma_\mathrm{i}\sqrt{\pi}}e^{-\frac{x_i^2}{\sigma_i^2}},\\&\int_{-\infty}^{\infty}dx_1\int_{-\infty}^{\infty}dx_2\ldots.\int_{-\infty}^{\infty}dx_\mathrm{N}P(x_1,x_2,\ldots x_\mathrm{N})=1.\end{aligned}\tag{2.19}$$

From (2.19) the contour of constant probability density is given by the N-dimensional ellipsoid surface with principal axes defined by the σ_i's

$$\sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2}=\mathrm{x}^{\mathrm{T}}\cdot\mathrm{E}\cdot\mathrm{x}=\mathrm{S}=-\ln\left(P\left(x_1,x_2,\ldots x_{\mathrm{N}}\right)\cdot\sqrt{\pi}^{\mathrm{N}}\prod_{\mathrm{i}}^{\mathrm{N}}\sigma_{\mathrm{i}}\right),$$
$$\mathrm{E}=\begin{pmatrix}1/\sigma_1^2 & 0 & 0 & 0\\ 0 & 1/\sigma_2^2 & 0 & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & 1/\sigma_{\mathrm{N}}^2\end{pmatrix}. \tag{2.20}$$

The diagonal matrix **E** has been encountered repeatedly as the "ellipsoid matrix" as in (2.8), and S is the "radius" of the ellipsoid surface. When the co-ordinates x_1, x_2, ...x_N are mapped into a different space spanned by basis y_1, y_2, ...y_M, not necessarily of the same dimensionality, by a matrix **M** with real coefficients, the ellipsoid can be re-scaled and sheared, namely **E** can develop off-diagonal terms and its magnitude will change. Nonetheless it remains symmetric and has only positive eigenvalues. It will be shown in Appendix I that under such a transformation, if one still demands the distribution of any of the new variables y_j in the image space with all other coordinates integrated out, the desired distribution will still be a Gaussian with a scaled σ_j determined purely by **M**, provided care has been taken to normalise the initial σ_j's,

$$P\left(y_{\mathrm{j}}=a\right)=\frac{1}{\sqrt{\pi}\left|\mathbf{M}_{\mathrm{j}}\right|}e^{-a^2/\left|\mathbf{M}_{\mathrm{j}}\right|^2}, \tag{2.21}$$

where $\left|\mathbf{M}_{\mathrm{j}}\right|$ stands for the length of the j-th row of **M**.

Some distributions of interest to the multi-dimensional Gaussian distribution concern those of the extremum or length. In Appendix I we show that for uncoupled and normalised distributions these can be evaluated in closed form as

$$P_{|max|}(\mathbf{m})=\frac{2\mathrm{NG}}{\sqrt{\pi}}e^{-\mathbf{m}^2}\mathrm{erf}^{\,\mathrm{N-1}}(\mathbf{m}) \tag{2.22}$$

for the distribution of absolute maximum and

$$P_{length}(\mathbf{r})=\frac{2}{\Gamma(\mathrm{N}/2)}\mathbf{r}^{\mathrm{N-1}}e^{-\mathbf{r}^2} \tag{2.23}$$

for the distribution of length. However, if the distribution shows correlation between the coordinates, the problem quickly becomes intractable and in fact there may be no known expression for the distribution of either the extremum or the length in closed form [8].

An efficient method, up to a few hundred dimensions, for obtaining the cutoff values at which the dominant portion of a length (or RMS) distribution is included, is developed using higher order inflection points of the distribution. This is discussed in Appendix I.

2.2.5 Singular Value Decomposition, Condition Number, Gram Determinant, Principal Axes, and Other Singularity Related Issues

Singular value decomposition (SVD) is the process of decomposing a matrix **M** into the product of three matrices **U**, **W** and **V**:

$$\mathbf{M} = \mathbf{U}^{\mathrm{T}} \cdot \mathbf{W} \cdot \mathbf{V}, \tag{2.24}$$

where **W** is a diagonal matrix with monotonically decreasing diagonal elements by convention, called singular values. The rows of **U** and **V** are orthonormal vectors in their respective dimensions. Useful information can be extracted from **U**, **W** and **V** when applied to response matrices. The rows of **V** represent combinations of the actuators, whose effects are magnified by the diagonal elements of **W** before being realised as responder patterns represented by the rows of **U**. SVD allows to decompose the response matrix into decoupled cause-effect relations between linear combinations of the actuators and responders with well defined magnification factors. The condition number of the matrix **M** can be defined as the ratio between the largest and the smallest singular values,

$$\mathbf{N}_{\mathrm{SVD}}^{\mathbf{M}} = \mathbf{W}_{11} / \mathbf{W}_{\mathrm{NN}}. \tag{2.25}$$

When a matrix **M** is near singular, it can be numerically “fixed” to prevent propagation of numerical instabilities to later calculations. This is done by decomposing **M** via (2.24), eliminating the rows of **U** and **V** and diagonal elements of **W** corresponding to singular values smaller than a pre-set tolerance, then recombining them to get a non-singular matrix which is close to **M** but without the near singular responses of **M**.

The Gram determinant $\mathbf{G_M}$ of a matrix **M** is given by

$$\begin{aligned} \mathbf{G_M} &= \mathrm{Det}\left(\mathbf{M}^{\mathrm{T}} \cdot \mathbf{M}\right) && \mathrm{Nr} \geq \mathrm{Nc}, \\ \mathbf{G_M} &= \mathrm{Det}\left(\mathbf{M} \cdot \mathbf{M}^{\mathrm{T}}\right) && \mathrm{Nr} \leq \mathrm{Nc}, \\ \mathbf{G_M} &= \left(\prod_{\mathbf{j}} \mathrm{S}_{\mathbf{j}}^{\mathbf{M}}\right)^2, \end{aligned} \tag{2.26}$$

where Nr and Nc are the number of rows and columns of **M**, and $\mathrm{S_j}$ is the j-th singular value of **M**. Being a measure of the “volume” of the matrix **M**, the Gram determinant can be used as an objective cutoff value for small singular values as

$$\mathrm{V}_{\mathrm{cutoff}} = \left|\mathbf{G_M}\right|^{\frac{1}{\mathrm{N}}}, \tag{2.27}$$

where N is the smaller dimension of **M**.

Given a symmetric ellipsoid represented by a square symmetric matrix **E**, the lengths and directions of its principal axes are simply given by the diagonal elements of **W** and row vectors of **V** in (2.24) when an SVD is applied to **E**.

2.3 Secondary Generalised Response Matrices

In a sense any process of orbit perturbation or manipulation, even the idiosyncratic way in which an operator performs orbit correction, when taken to the first order can be expressed as a response matrix. The case of the orbit correction process is significant in that it involves a one-step feedback[2], namely, the responder signal is used as the input back into the process. In this case an extra set of errors, those due to the limited accuracy of the monitors, comes into play. It is indistinguishable from the other actuators of Table 1 for the purpose of formulating the problem. The only distinction is in their response coefficients and in their very different effects on orbit correction [5].

We can thus construct more "response matrices" capturing the orbit correction process under various premises. First we need to construct the trivial matrix $\mathbf{M}^{\mathbf{MM}}$ linking the monitor offset error to the apparent error at the monitor due to the error $\mathbf{M}_{ij}^{\mathbf{MM}}=\delta_{ij}$, namely. $\mathbf{M}^{\mathbf{MM}}$ is simply the N×N square identity matrix where N is the total number of monitors. Despite its trivial form, this matrix closes the loop of the one-step feedback and enables us to formally represent the orbit correction process and incorporate the expression of the true "underlying" orbit error within the given framework, as will become evident in what follows.

Other response matrices are discussed in the following sub-sections. Almost all distributions discussed in this paper can trace their origins to the actuator spaces represented by errors, either alignment or monitor. The errors start as independent distributions with different RMS, or σ's. It would help the formulation considerably to re-scale the error units such that all σ's are equal to unity, as discussed in (I.1) of Appendix I. This will be the assumption for the remainder of the report. In particular the identity matrix $\mathbf{M}^{\mathbf{MM}}$ discussed above will take on diagonal entries other than 1 under this re-scaling.

2.3.1 Observability Matrix Purely due to Monitor Configuration

We define the matrices:

$$\begin{aligned}
\mathbf{M}_{\mathbf{M}}^{\text{unobs}} &= \mathbf{M}^{\mathbf{EA}} \cdot \left(\mathbf{M}^{\mathbf{EM}}\right)_{\text{out}}^{\mathbf{M}^{\mathbf{EM}}}, \\
\mathbf{M}_{\mathbf{M}}^{\text{obs}} &= \mathbf{M}^{\mathbf{EA}} \cdot \left(\left(\mathbf{M}^{\mathbf{EM}}\right)_{\text{in}}^{\mathbf{M}^{\mathbf{EM}}}\right)^{-1} \mathbf{M}^{\mathbf{MM}}.
\end{aligned} \tag{2.28}$$

The definition of the decoupled matrix for matrix $\mathbf{M}^{\mathbf{EM}}$ is as given in (C.2). The significance of the first matrix is that it maps from the (orthonormalised) sub-space of the error space that lies outside the monitoring power of the monitors onto the all-element space. The second matrix maps from the monitor space through the observable part of the error space onto the all-element space, and thus represents an effect on the underlying orbit for a given observed orbit. The matrix $\mathbf{M}^{\mathbf{MM}}$ is formally the same as discussed earlier for representing monitor errors, but here takes on a different meaning of representing simply monitor readings. One should bear in mind that it is not necessarily the identity matrix after the normalisation.

Because of the a priori normalisation of all axes to make the source distributions have σ's of 1, the orthonormalisation involved in (2.28) does not necessitate a re-examination of the source distribution.

An alternative way to view the underlying orbit is to place the monitor offset error and the alignment error on the same probabilistic footing by forming the direct-sum space of $\mathbf{A}^{\mathbf{E}}$ and $\mathbf{A}^{\mathbf{M}}$:

$$\begin{aligned}
&\mathbf{M}_{\mathbf{M}}^{\text{DS-unobs}} = \mathbf{K}_{\mathbf{A}} \cdot \Pi_{\mathbf{K}_{\mathbf{M}}}^{\perp}, \\
&\mathbf{K}_{\mathbf{A}} = \left(\mathbf{M}^{\mathbf{EA}} \oplus \mathbf{Z}^{\mathbf{MM}}\right), \\
&\mathbf{Z}^{\mathbf{MM}} = \mathrm{Z}_{ij}^{\mathrm{MM}} = 0,\ \mathrm{i}=1,2,\ldots.\mathrm{N}_{\mathrm{A}},\ \mathrm{j}=1,2,\ldots.\mathrm{N}_{\mathrm{M}}, \\
&\mathbf{K}_{\mathbf{M}} = \left(\mathbf{M}^{\mathbf{EM}} \oplus \mathbf{M}^{\mathbf{MM}}\right).
\end{aligned} \tag{2.29}$$

[2] As model error is not considered here, repetition of this feedback process is meaningless.

The meaning of (2.29) is as follows. Assume that there is a configuration of alignment and monitor offset errors under a combined probability distribution. The quantity $\mathbf{M}_{\mathbf{M}}^{\text{DS-unobs}}$ projects out the unobservable part of this configuration and maps it onto all elements, namely, the component of the combined configuration inside the range of $\mathbf{M}_{\mathbf{M}}^{\text{DS-unobs}}$ will cause all monitors to read 0. Assume further that the physical setting is maintained, but that the monitors are re-aligned such that they all have 0 offset, then the orbit displayed at all monitors would be exactly equal to (minus) the original monitor offsets. These two equivalent pictures are both described by $\mathbf{M}_{\mathbf{M}}^{\text{DS-unobs}}$, which links the combined error space and the underlying orbit errors at all elements. As opposed to treating the alignment and monitor offset errors separately, this view puts them on equal probabilistic footing.

2.3.2 Corrector Response Matrix to Errors, Ultimate Limit on Correcting Power

We define the matrix

$$\mathbf{M}_{\mathbf{CME}}^{\text{resp}} = \mathbf{M}_{\mathbf{CM}}^{\dagger} \cdot \mathbf{M}^{\mathbf{EM}}. \tag{2.30}$$

This is understood as the mapping from the error space to the corrector space following an SVD-type correction. The SVD scheme is adopted both because it reflects the fundamental division between correctable and not correctable spaces, a property inescapable for any correction scheme, and because it yields the minimal RMS solution in the under-constrained case, a favourable option for the analysis of corrector limits.

We can further define

$$\mathbf{M}_{\mathbf{CAE}}^{\text{resp}} = \mathbf{M}_{\mathbf{CA}}^{\dagger} \cdot \mathbf{M}^{\mathbf{EA}}. \tag{2.31}$$

This has the same meaning as (2.30), but assuming unlimited monitoring power in the entire line, thus is characteristic of the ultimate correction ability given the corrector configuration, independent of that of the monitors.

2.3.3 Correctable / Uncorrectable Errors at Monitors and All-Elements and Orbit Errors Induced by Monitor Offsets

We define the following matrices

$$\begin{aligned}
\mathbf{M}_{\mathbf{CM}}^{\text{uncorr}} &= \Pi_{\mathbf{CM}}^{\perp} \cdot \mathbf{M}^{\mathbf{EM}}, \\
\mathbf{M}_{\mathbf{CA}}^{\text{uncorr}} &= \Pi_{\mathbf{CA}}^{\perp} \cdot \mathbf{M}^{\mathbf{EA}}, \\
\mathbf{M}_{\mathbf{CM}}^{\text{corr}} &= \Pi_{\mathbf{CM}}^{\parallel} \cdot \mathbf{M}^{\mathbf{MM}}.
\end{aligned} \tag{2.32}$$

The first matrix projects out the error space that is beyond correction with the given monitor and corrector configuration. The second matrix bears the same interpretation but with unlimited monitoring power, similar to (2.31). The third matrix projects out the monitor error space that is correctable in the given monitor and corrector configuration, leading to undetected steering errors.

We can also establish

$$\mathbf{M}_{\mathbf{MCA}}^{\text{ind}} = \mathbf{M}^{\mathbf{CA}} \cdot \mathbf{M}_{\mathbf{CM}}^{\dagger} \cdot \mathbf{M}^{\mathbf{MM}} \tag{2.33}$$

as defining the orbit error at all elements induced by monitor offsets. Monitor offsets contribute a distinct component to the eventual orbit error through correction [5]. This error depends on the configurations of both the monitors and correctors.

2.3.4 Residue of Orbit Correction

We define the matrices

$$\begin{aligned}\mathbf{M}^{\text{resid}}_{\mathbf{CMA}} &= \mathbf{M}^{\mathbf{EA}} - \mathbf{M}^{\mathbf{CA}} \cdot \mathbf{M}^{\text{resp}}_{\mathbf{CM}} = \mathbf{M}^{\mathbf{EA}} - \mathbf{M}^{\mathbf{CA}} \cdot \mathbf{M}^{\dagger}_{\mathbf{CM}} \cdot \mathbf{M}^{\mathbf{EM}}, \\ \mathbf{M}^{\text{resid}}_{\mathbf{CAA}} &= \mathbf{M}^{\mathbf{EA}} - \mathbf{M}^{\mathbf{CA}} \cdot \mathbf{M}^{\text{resp}}_{\mathbf{CA}} = \mathbf{M}^{\mathbf{EA}} - \mathbf{M}^{\mathbf{CA}} \cdot \mathbf{M}^{\dagger}_{\mathbf{CA}} \cdot \mathbf{M}^{\mathbf{EA}}.\end{aligned} \tag{2.34}$$

These directly link the errors to the residue orbit after steering with the given corrector configuration. The first matrix assumes the given monitor configuration, and the second one assumes unlimited monitoring power, thus providing a measure of the ultimate limit to the corrector configuration.

The residue orbit at all monitors is linked to the errors through

$$\mathbf{M}^{\text{resid}}_{\mathbf{CMM}} = \mathbf{M}^{\mathbf{EM}} - \mathbf{M}^{\mathbf{CM}} \cdot \mathbf{M}^{\text{resp}}_{\mathbf{CM}} = \mathbf{M}^{\mathbf{EM}} - \mathbf{M}^{\mathbf{CM}} \cdot \mathbf{M}^{\dagger}_{\mathbf{CM}} \cdot \mathbf{M}^{\mathbf{EM}}. \tag{2.35}$$

2.3.5 Implication on Errors and Correctors from Residual Orbit

Assume the numbers of errors, monitors and correctors are N_E, N_M, and N_C, respectively, we define the matrix

$$\begin{aligned}\mathbf{M}^{\text{rms-orb}}_{\mathbf{EC}} &= \left(\mathbf{M}'^{\mathbf{EA}} \cdot \mathbf{P}_E + \mathbf{M}^{\mathbf{CA}} \cdot \mathbf{P}_C\right) \cdot \mathbf{M}^{\mathbf{MM}} , \\ \mathbf{P} = \left(\mathbf{M}'^{\mathbf{EM}} \oplus \mathbf{M}^{\mathbf{CM}}\right)^{\dagger} &= \left(\begin{array}{ccc} P^{11} & \cdots & P^{1,N_M} \\ \vdots & \ddots & \vdots \\ P^{N_E,1} & \cdots & P^{N_E,N_M} \\ \hline P^{N_E+1,1} & \cdots & P^{N_E+1,N_M} \\ \vdots & \ddots & \vdots \\ P^{N_E+N_C,1} & \cdots & P^{N_E+N_C,N_M} \end{array}\right) = \begin{pmatrix} \mathbf{P}_E \\ \mathbf{P}_C \end{pmatrix}.\end{aligned} \tag{2.36}$$

This matrix maps a given residual orbit pattern at the monitors, via a minimal error-corrector combination, onto the corresponding residual orbit pattern at all elements. The matrix $\mathbf{M'}^{\mathbf{EA}}$ and $\mathbf{M'}^{\mathbf{EM}}$ differ from $\mathbf{M}^{\mathbf{EA}}$ and $\mathbf{M}^{\mathbf{EM}}$ in that the former do not go through the normalisation making all σ's unity. This is because the matrix combined under the direct sum is underdetermined; normalisation of only $\mathbf{M}^{\mathbf{EM}}$ without scaling $\mathbf{M}^{\mathbf{CM}}$ accordingly would distort the pseudo-inverse.

Two other points worth noticing are: firstly, in using (2.36) one normally starts with a given RMS in the residual orbit, as opposed to total orbit magnitude more relevant in other problems. There is thus an enhancement factor of $\sqrt{N_M}$ associated with this part of the analysis. Secondly, the final magnitude of the residual orbit at all elements should be limited by the error magnitude through which the original residual orbit at the monitors is realised. Thus one should first convert the derived residual orbits into corresponding error RMS and scale down the former if necessary until the latter is within a reasonable range. One should however bear in mind that for multi-dimensional error distributions, the "reasonable" cutoff for the error RMS can be very large, due to the shifted distribution peak discussed in Appendix I. A well-defined method for defining this cutoff for a given N_E is given there.

2.3.6 Actual Underlying Orbit Error Including Effects from Monitor Offsets

In the presence of monitor errors, one needs to introduce a new "actuator" and two new "responders". In a probabilistic picture the probability density distribution for the combined alignment and monitor errors should simply be that obtained by taking all these errors as independent, preferably with normalisation. Thus a combined error actuator $\mathbf{A}^{\mathbf{EM}}$ can be constructed as (seeTable 1)

$$\mathbf{A}^{\mathbf{EM}} = \mathbf{A}^{\mathbf{E}} \oplus \mathbf{A}^{\mathbf{M}} = \left(A_1^E, \ldots A_{N_E}^E, A_1^M, \ldots A_{N_M}^M\right)^T . \tag{2.37}$$

On the other hand, the underlying orbit error which, as opposed to the apparent orbit error $\mathbf{R}^{\mathbf{M}}$ observed on monitors, represents the real orbit error with respect to the design baseline, even if the apparent orbit errors are all 0. The distinction between these two orbit errors comes into being, of course, due to monitor offset errors that compromise orbit correction effectiveness [5]. One can denote this $\mathbf{R}^{\mathbf{UM}}$ by:

$$\mathbf{R}^{\mathbf{UM}} = \Pi_{\mathbf{CM}}^{\perp} \cdot \mathbf{M}^{\mathbf{EM}} \cdot \mathbf{A}^{\mathbf{E}} - \Pi_{\mathbf{CM}}^{\parallel} \cdot \mathbf{M}^{\mathbf{MM}} \cdot \mathbf{A}^{\mathbf{M}}, \tag{2.38}$$

These quantities would not be useful if they could not be linked through a single matrix. This can be done through

$$\mathbf{M}_{\mathbf{EMM}}^{\mathrm{UOE}} = \left(\Pi_{\mathbf{CM}}^{\perp} \cdot \mathbf{M}^{\mathbf{EM}}\right) \oplus \left(-\Pi_{\mathbf{CM}}^{\parallel} \cdot \mathbf{M}^{\mathbf{MM}}\right), \tag{2.39}$$

where the direct sum (⊕) simply concatenates the two matrices column-wise.

The minus signs in (2.38) and (2.39) do not affect the answer to questions about maximal error in underlying orbit using ellipsoidal projection methods, since the ellipsoid is symmetric. They do however keep the correct relative sign between the two terms [5], which is important when one wants to derive the error combinations which cause a particular type of errors.

A more elaborate but also more useful responder is the actual underlying orbit error $\mathbf{R}^{\mathbf{UA}}$ at all elements, and the matrix $\mathbf{M}_{\mathbf{EMA}}^{UOE}$ to link it to $\mathbf{A}^{\mathbf{EM}}$:

$$\begin{aligned} &\mathbf{M}_{\mathbf{EMA}}^{\mathrm{UOE}} = \left(\mathbf{M}^{\mathbf{EA}} \oplus \mathbf{Z}^{\mathbf{MM}}\right) - \mathbf{M}^{\mathbf{CA}} \cdot \mathbf{M}_{\mathbf{CM}}^{\dagger} \cdot \left(\mathbf{M}^{\mathbf{EM}} \oplus \mathbf{M}^{\mathbf{MM}}\right), \\ &\mathbf{R}^{\mathbf{UA}} = \mathbf{M}_{\mathbf{EMA}}^{\mathrm{UOE}} \cdot \mathbf{A}^{\mathbf{EM}}, \\ &\mathbf{Z}^{\mathbf{MM}} = Z_{ij}^{MM} = 0, \ i=1,2,\ldots.N_A, \ j=1,2,\ldots.N_M . \end{aligned} \tag{2.40}$$

2.3.7 Summary of Secondary Response Matrices

Table 2 summarises the secondary response matrices established so far along with the spaces they act on and brief descriptions:

Response Matrix	Domain	Image Space	Physical Significance
$\mathbf{M}_{\mathbf{M}}^{\mathrm{unobs}}$	$\mathbf{A}^{\mathbf{E}}$	$\mathbf{R}^{\mathbf{A}}$	Unobservable error
$\mathbf{M}_{\mathbf{M}}^{\mathrm{obs}}$	$\mathbf{A}^{\mathbf{M}}$	$\mathbf{R}^{\mathbf{A}}$	Effect of observed orbit
$\mathbf{M}_{\mathbf{M}}^{\mathrm{DS\text{-}unobs}}$	$\mathbf{A}^{\mathbf{EM}}$	$\mathbf{R}^{\mathbf{A}}$	Underlying orbit at all elements under given monitor orbit
$\mathbf{M}_{\mathbf{CME}}^{\mathrm{resp}}$	$\mathbf{A}^{\mathbf{E}}$	$\mathbf{A}^{\mathbf{C}}$	Corrector strength needed for error
$\mathbf{M}_{\mathbf{CAE}}^{\mathrm{resp}}$	$\mathbf{A}^{\mathbf{E}}$	$\mathbf{A}^{\mathbf{C}}$	Corrector strength needed (unlimited monitor power)
$\mathbf{M}_{\mathbf{CM}}^{\mathrm{uncorr}}$	$\mathbf{A}^{\mathbf{E}}$	$\mathbf{A}^{\mathbf{E}}$	Uncorrectable error
$\mathbf{M}_{\mathbf{CA}}^{\mathrm{uncorr}}$	$\mathbf{A}^{\mathbf{E}}$	$\mathbf{A}^{\mathbf{E}}$	Uncorrectable error (unlimited monitor power)
$\mathbf{M}_{\mathbf{CM}}^{\mathrm{corr}}$	$\mathbf{A}^{\mathbf{M}}$	$\mathbf{A}^{\mathbf{M}}$	Correctable monitored orbit or monitor error
$\mathbf{M}_{\mathbf{MCA}}^{\mathrm{ind}}$	$\mathbf{A}^{\mathbf{E}}$	$\mathbf{R}^{\mathbf{A}}$	Monitor error induced orbit at all elements
$\mathbf{M}_{\mathbf{CMA}}^{\mathrm{resid}}$	$\mathbf{A}^{\mathbf{E}}$	$\mathbf{R}^{\mathbf{A}}$	Residual orbit at all elements after correction at monitors
$\mathbf{M}_{\mathbf{CAA}}^{\mathrm{resid}}$	$\mathbf{A}^{\mathbf{E}}$	$\mathbf{R}^{\mathbf{A}}$	Residual orbit at all elements (unlimited monitor power)
$\mathbf{M}_{\mathbf{CMM}}^{\mathrm{resid}}$	$\mathbf{A}^{\mathbf{E}}$	$\mathbf{R}^{\mathbf{M}}$	Residual orbit at monitors after correction
$\mathbf{M}_{\mathbf{EC}}^{\mathrm{rms\text{-}orb}}$	$\mathbf{A}^{\mathbf{M}}$	$\mathbf{R}^{\mathbf{A}}$	Effect implied by observed orbit on error and corrector
$\mathbf{M}_{\mathbf{EMM}}^{\mathrm{UOE}}$	$\mathbf{A}^{\mathbf{EM}}$	$\mathbf{R}^{\mathbf{UM}}$	Real underlying orbit error at monitors
$\mathbf{M}_{\mathbf{EMA}}^{\mathrm{UOE}}$	$\mathbf{A}^{\mathbf{EM}}$	$\mathbf{R}^{\mathbf{UA}}$	Real underlying orbit error at all elements

Table 2: Secondary response matrices

2.4 Setting up the Actuator and Responder Arrays for the Transfer Lines

In this section we describe the strategy followed to establish the content of the actuator and responder arrays. This amounts to the judicial selection of representative elements in the beam line, determination of relevant error sources and magnitudes, and accurate representation of the orbit correction elements involved, including possibly non-physical ones.

2.4.1 Setting up the All-Element Actuator and Responder Arrays

For a representative set of elements serving the purpose of monitoring orbit responses throughout the beam line, we adopted the following principles in the analysis program:

- Adjacent elements should not be separated by more than a user specified amount in betatron phase.
- Adjacent elements should not be separated by more than a user specified amount in physical distance.
- Monitors can be eliminated from the list to avoid certain artificial degeneracy in some algorithms.
- Trajectory angle at the exit of the beam line can be included as a special monitoring entry.

Of the above, the element set created by the first two criteria can be combined either through a UNION or an INTERSECTION, determined by the user. For the lines analysed typical inter-element increments of 1 meter in distance and 3.5° in betatron phase were used, as were the INTERSECTION option and exclusion of monitors. This resulted in a quite dense coverage of the TI 2 line, for example, with about 350 elements out of a possible maximum of 1000. For a safety check this was compared to a scheme using increments of 1 meter in distance and 2.0° in betatron phase, the UNION option and inclusion of monitors, resulting in a very dense coverage of the TI 2 line with about 750 elements. No difference was observed in all cases tested.

The exit angle element is important in revealing possible flaws in an orbit correction system having negative impact on the exit angle of a beam line. This is not observable from the position alone at any physical location.

Another all-element array was established as actuators representing potential candidates for new correctors. This was simply taken to be the entire collection of beam line elements.

2.4.2 Setting up the Error Actuator Array

The source identification and magnitude assessment of errors largely follow the line of reasoning presented in a previous analysis [2], with the notable exception of the addition of injection position and angle errors and a conversion from a square distribution of the monitor offset error to Gaussian. They are briefly listed in the following.

- Injection position and angle errors: These are taken to be Gaussian distributions with an injection position σ of 0.5 mm and an injection angle σ of 0.05 mrad. Other values have been used for studies focused on constraints imposed by the injection condition.

- Dipole field errors: These are taken to be Gaussian with a σ corresponding to 0.000167 of the individual bending field. This value is selected such that when the distribution is cut off at 3 σ, the extent of the error magnitude, which is the focus of this study, is exactly equal to that of the distribution with the RMS (0.00025 of the field) and cutoff (2 σ) adopted in [2]. Further processing is needed to account for the finite length effect of a dipole field error. For this the end position effect was included

  $$\delta P = \delta A \cdot L^2/2 \tag{2.41}$$

  where δP, δA and L are the total end position error, total end angle error, and total dipole length, respectively. The response matrix element from the end of a dipole **D** to a downstream point **A** due to this combined error is then

  $$\begin{aligned} \mathrm{M}_{\mathrm{DA}}^{1} &= \mathrm{M}_{\mathrm{DA}}^{11} \cdot \delta P + \mathrm{M}_{\mathrm{DA}}^{12} \cdot \delta A = \left(\mathrm{M}_{\mathrm{DA}}^{11} \cdot L^2/2 + \mathrm{M}_{\mathrm{DA}}^{12}\right) \cdot \delta A, \\ \mathrm{M}_{\mathrm{DA}}^{2} &= \mathrm{M}_{\mathrm{DA}}^{21} \cdot \delta P + \mathrm{M}_{\mathrm{DA}}^{22} \cdot \delta A = \left(\mathrm{M}_{\mathrm{DA}}^{21} \cdot L^2/2 + \mathrm{M}_{\mathrm{DA}}^{22}\right) \cdot \delta A, \end{aligned} \tag{2.42}$$

  representing the position and angle responses at **A** to δA at **D**.

- Dipole tilt error: These are taken to be Gaussian with a σ corresponding to 0.000267 radian. This value is selected such that when the distribution is cut off at 3 σ, the extent of the error magnitude, which is the focus of this study, is exactly equal to that of the distribution with the RMS (0.0002 of the field) and cutoff (4 σ) adopted in [2]. The same processing accounting for finite dipole length as (2.42) is implemented.

- Quadrupole offset: These are taken to be Gaussian with a σ corresponding to 0.0002 m and a 3 σ cutoff, again identical to the values adopted in [2]. The kick produced at each quadrupole due to this offset is then calculated based on the quadrupole strength.

- Monitor offset: These are taken to be Gaussian with a σ corresponding to 0.000288675 m. In [2] it was taken to be a rectangular distribution cut off at ±0.0005 m, which amounted to the same RMS as the Gaussian used here. We choose to align the RMS value with the original study since in the latter the subtle effect of the square distribution in the mixed distributions should not be easily dismissed, and it is best accounted by using its RMS. This resulted in a 3 σ cutoff of the Gaussian distribution at 0.000866 m, which is a little pessimistic compared to the extent of the square distribution at 0.0005 m, but it should more faithfully reflect the probabilistic impact of the latter.

2.4.3 Setting up the Monitor Actuator and Responder Arrays

The monitor arrays as actuators and responders are established based on the monitor configuration under study. As responders they are the position and/or angle at the monitors, and as actuators they are the monitor offsets or resolution at the monitors.

The horizontal and vertical planes are not assumed to have the same monitor configuration, as single-plane monitor is a valid option, allowing taking advantage of different local lattice properties in the two planes at minimal cost.

A special “end angle” monitor can be added in case one wants to mimic the effect of position monitors beyond the beam line being studied. Such monitors have a constraining effect on the outgoing angle from the line which, if not taken into account, can lead to an artificially large orbit envelope at the end.

The monitor configurations for TI 2 and TI 8 are extended into the injection sections for the LHC to accurately reflect the performance at the end of these lines. At least two monitors beyond the end of TI 2 and TI 8 are required to provide adequate (and correct) orbit anchoring, or an artificially large orbit envelope would develop towards the end. In the current analysis this potential artefact is rectified by extending the lines to encompass one downstream monitor in the LHC injection section, and an angle monitor at the end of the entire configuration, to represent the proper orbit-anchoring beyond the TI 2 and TI 8 lines.

The monitor array as actuators are used in two different contexts in the analysis, as monitor offsets and resolutions, with the latter characterising the precision limit of the monitors. The response matrices acting on these two entities are of course identical, but the error envelopes are assigned different σ’s for different interpretations.

2.4.4 Setting up the Corrector Actuator Array

The corrector arrays as actuators are established based on the corrector configuration under study. Further processing is needed in two cases:

- For dipoles as correctors the finite-length effect of (2.42) must be implemented.
- For correctors/dipoles powered in series their combined effect is entered as a single entry in the corrector array.

For realistic systems all correctors are limited in range. For the analysis in this report the ranges are defined as follows:

- For standard dedicated orbit correctors a limit of ±80 μrad is used [3].

- For dipoles or dipole strings a range of ±120 μrad available for orbit correction is assumed, in addition to what is needed to define the baseline geometry.

2.4.5 Establishing the Primary and Secondary Response Matrices

Once the actuator and responder arrays are identified, the primary response matrices can be established from the optical transfer properties of the line. This can involve all four transfer matrix elements in each plane, not only $\mathbf{M}_{12}$ and $\mathbf{M}_{34}$, as explained in the previous sub-sections. The secondary transfer matrices follow straightforwardly from the primary ones.

2.5 Evaluating the Performance of the Orbit Correction Configuration

Equipped with the tools developed so far, we can begin to examine the performance of a given orbit correction system, and improve it in an analytical manner if necessary. The main theme for what remains of this report is to apply the techniques developed in Section 2.2 and the Appendices concerning the response matrices of Table 1 and Table 2, and extract information about the performance of an orbit correction system. As explanation of the application of these techniques to real physical systems proceeds, examples will be freely drawn from studies which where made on the transfer lines TI 2 and TI 8. If not noted otherwise, numerical values used in the examples, such as σ's for the quadrupole offset distribution or corrector limits, are as defined in the previous section. Definition of all primary and secondary response matrices can be found in Sections 2.1 and 2.3.

2.5.1 Ellipsoidal Surface of Constant Probability Density in the Actuator Space

The predominant part of the following analysis will start with the constant probability ellipsoidal surface defined in (2.20)[3]. For most of the applications it is further normalised as in (I.1). Thus, for example, if we start with the complete distribution in the error space $\mathbf{A}^{\mathbf{E}}$**,** the re-scaling matrix C_E normalises the distribution so that the constant probability density contours are spheres, and $\mathbf{A}^{\mathbf{E}}$ and the response matrices $\mathbf{M}^{\mathbf{EM}}$ and $\mathbf{M}^{\mathbf{EA}}$ will undergo transformations:

$$C_E = \begin{pmatrix} 1/\sigma_1^E & 0 & 0 & 0 \\ 0 & 1/\sigma_2^E & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1/\sigma_N^E \end{pmatrix},$$

$$\mathbf{A}^{\mathbf{E}} \rightarrow C_E \cdot \mathbf{A}^{\mathbf{E}},\ \mathbf{M}^{\mathbf{EM}} \rightarrow \mathbf{M}^{\mathbf{EM}} \cdot C_E^{-1},\ \mathbf{M}^{\mathbf{EA}} \rightarrow \mathbf{M}^{\mathbf{EA}} \cdot C_E^{-1}. \qquad (2.43)$$

This normalised ellipsoidal (spherical) surface then is mapped by various primary and secondary matrices according to the rules described in the Appendices to higher or lower dimensional spaces. Information about tangent points, extrema of various operators, and singularity is extracted from these projected surfaces based on techniques developed in Appendices D, E, F, G and H, while their physical implications are studied.

[3] Although we use the Gaussian example in this equation, we are not limiting the distribution to Gaussian for the validity of the discussion in this section, but only to those satisfying conditions mentioned in section 2.2.2, as long as the constant probability surfaces can be re-scaled to spheres.

Throughout the remainder of Section 2 graphical examples will be used to illustrate various performance criteria of an orbit correction system examined by this program. All examples will be taken from an intermediate configuration in the TI 2 line during the course of optimisation.

2.5.2 Fundamental Observability, Monitor Coverage, and Effects of Residual Orbit

Given a certain distribution of errors and monitor configuration in a beam line, an important question is whether the monitors can discern effects due to such errors. This question is answered by the following procedures.

2.5.2.1 Unobservable Error

Taking the matrix $\mathbf{M}_{\mathbf{M}}^{\text{unobs}}$ and the 3 σ error ellipsoid in its domain space (a subspace of $\mathbf{A}^{\mathbf{E}}$ by (2.28)), we can find the boundary of its corresponding image in $\mathbf{R}^{\mathbf{A}}$. This gives the orbit error unobservable by the monitors at an extent of 3 σ at all (representative) elements along the line. An example is depicted in Figure 3.

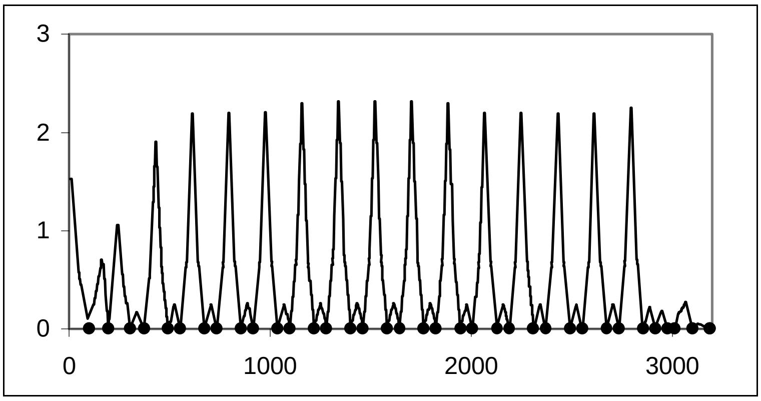


Figure 3: Full line: maximally unobservable orbit error in [mm] along the beam line in [m]; dots: position of monitors.

2.5.2.2 Effect of Finite Residual RMS Orbit

Taking the matrix $\mathbf{M}_{\mathbf{M}}^{\text{obs}}$ and the 1 σ error ellipsoid in its domain space (a subspace of $\mathbf{A}^{\mathbf{M}}$ by (2.28)) enhanced by a factor of $\sqrt{N_M}$, we can find the boundary of its corresponding image in $\mathbf{R}^{\mathbf{A}}$. This gives the maximal orbit error at all (representative) elements along the line, if the RMS residual orbit observed at the monitors is equal to the RMS value of the assumed monitor offsets[4]. This value should be further scaled at every location, so that the corresponding error magnitude is within a criterion determined by the cutoff (described in Appendix I) proper for the N-dimensional space. This is depicted in Figure 4.

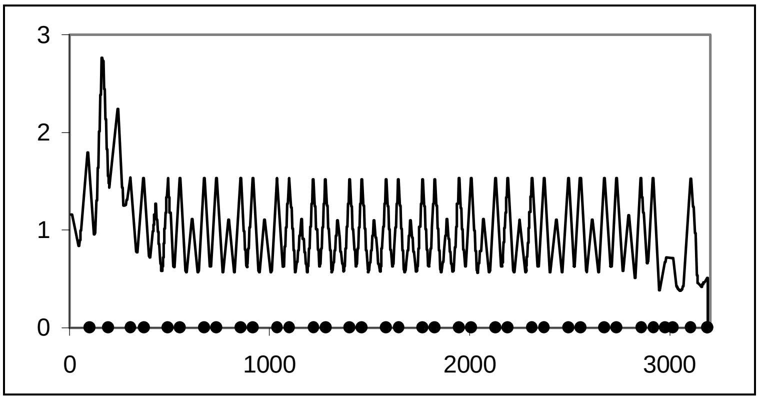


Figure 4: Full line: maximal orbit error in [mm] due to residual observed orbit along the beam line in [m]; dots: position of monitors.

2.5.2.3 Combined Effects of Finite Residual RMS Orbit

The orbit envelopes from the two previous sections come from

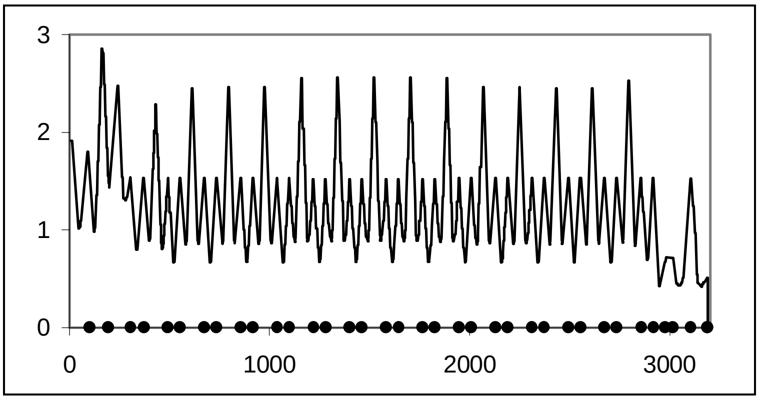


Figure 5: Full line: maximal orbit error in [mm] along the beam line in [m] (quadratic sum of Figure 3 and Figure 4 scaled by 3σ error); dots: position of monitors.

[4] This may be an over-pessimistic value both due to the RMS enhancement and the use of a standard monitor offset σ for the observed orbit. Using the assumed monitor resolution, which is smaller by a factor of 5, is an alternative.

completely orthogonal subspaces in the error space. We can take the quadratic sum of these two envelopes to obtain the real underlying orbit envelope at the same extent of probability density as individual ones. This is shown in Figure 5.

2.5.2.4 Effects of the Finite Residual RMS Orbit (Alternative View)

As explained in (2.29), an alternative view treating alignment and monitored orbit errors on an equal footing using $\mathbf{M}_{\mathbf{M}}^{\text{DS-unobs}}$ generates a different view on the underlying orbit as a function of the "total length" of the apparent orbit. This is shown in Figure 6. It is similar to the previous outcome, but slightly more optimistic, as expected.

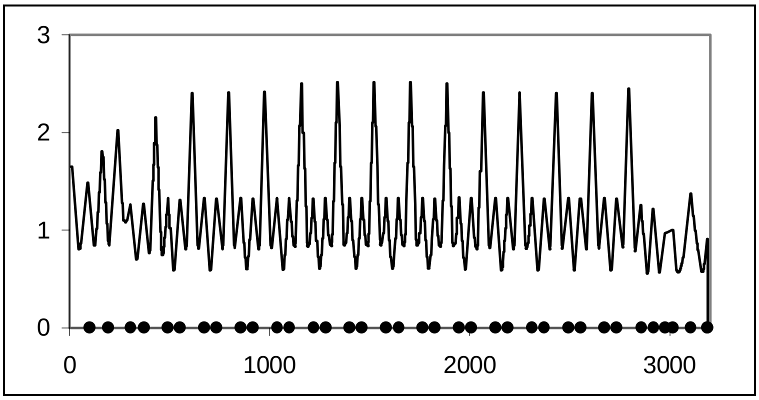


Figure 6: Full line: maximal orbit error (alternative view) in [mm] along the beam line in [m]; dots: position of monitors.

2.5.3 Monitor Efficiency and Redundancy

2.5.3.1 Monitor Efficiency

The efficiency of a monitor can be measured by the projection of the error ellipsoid $\mathbf{E}^{\mathbf{E}}$ as defined in (2.8) along all axes in the monitor space, using the method described in Appendix G. An example is given in Figure 7. Naturally, as the errors accumulate along a beam line, monitors will pick up progressively larger projections from the ellipsoid. Monitors located where the projection is significantly below this trend are considered to have below-average efficiency[5]. A plot such as in Figure 7 also reveals strategic locations and gaps in the monitor configuration.

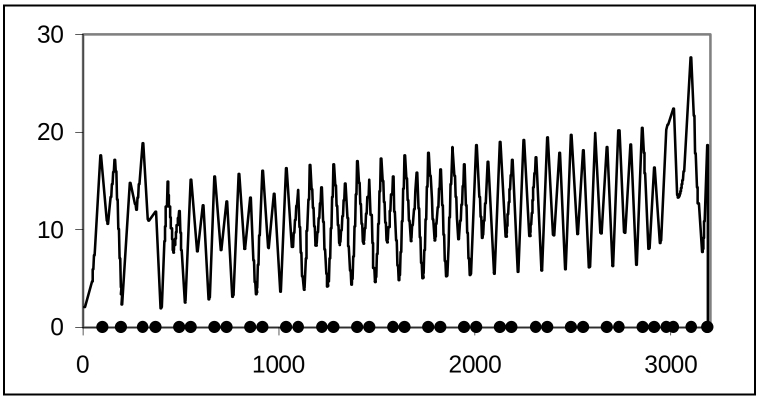


Figure 7: Full line: projected 3σ error onto orbit space in [mm] along the beam line in [m]; dots: position of monitors.

2.5.3.2 Redundant Monitor Combination

Another test of the degree of redundancy of the monitors, purely from the standpoint of cost

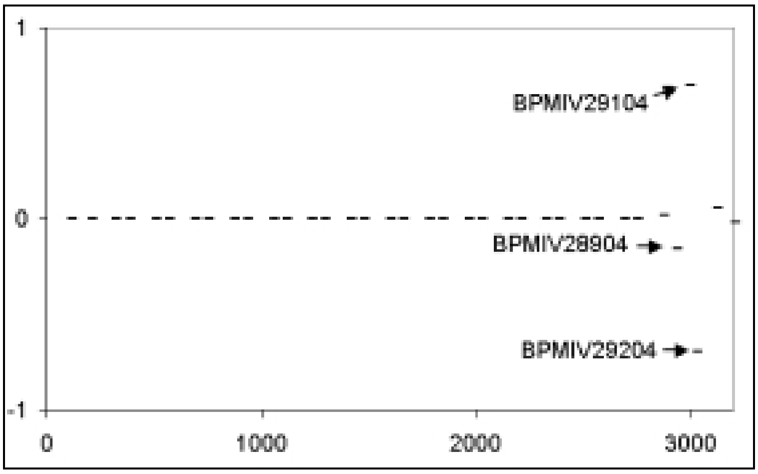


Figure 8: Example of orbit pattern with very small singular values, with total magnitude normalised to 1, along the beam line in [m].

[5] Of course this is not necessarily a criterion for the removal or relocation of this monitor. Special steering concerns and cost effectiveness also should be considered, as from a pure orbit-control point of view, one could never have too many monitors.

effectiveness, is to perform an SVD on the response matrix $\mathbf{M}^{\mathbf{EM}}$ as in (2.24), and examine the row vectors of $\mathbf{U}$ corresponding to singular values smaller than the cutoff defined by (2.27). These vectors will contain either one dominant monitor, indicating below-average efficiency as discussed above, or a combination of a few monitors, indicating an inefficient combination of monitors leading to unusable monitoring degrees of freedom. An iterative algorithm [5] can be applied to this configuration to ensure eventually the elimination of unwanted redundancy. An example output of the algorithm is shown in Figure 8 indicating a near singular combination of two monitors. Elimination of the dominant monitor of this pair did not compromise the monitoring power.

2.5.4 Fundamental Correctability, Correction Range and Residual Orbit

After having examined issues of observability and monitor efficiency, we need now to examine the effectiveness of the corrector configuration[6].

2.5.4.1 Corrector Range

By taking the projection of the error ellipsoid $\mathbf{E}^{\mathbf{E}}$ onto the corrector space via the secondary response matrix $\mathbf{M}^{\text{resp}}_{\mathbf{CME}}$ of (2.30), we can derive the maximum error, in units of the global error distribution σ, that can be handled by each corrector. Conversely, for a fixed magnitude of the global error distribution, we can sort the correctors in order of increasing correction range, thereby identifying weak links in terms of hardware limits in the correctors. Figure 7 demonstrates such an ordering, with the triangles indicating the correction range for each corrector in units of the global error distribution σ, along the beam line TI 2. The dotted line at 3 σ serves as reference.

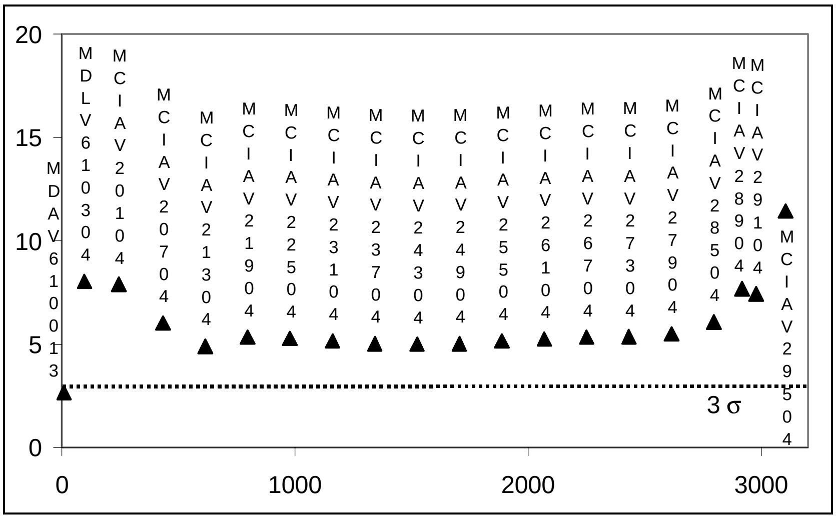


Figure 9: Triangles: correction range of correctors in [σ] of the global error distribution, along the beam line in [m].

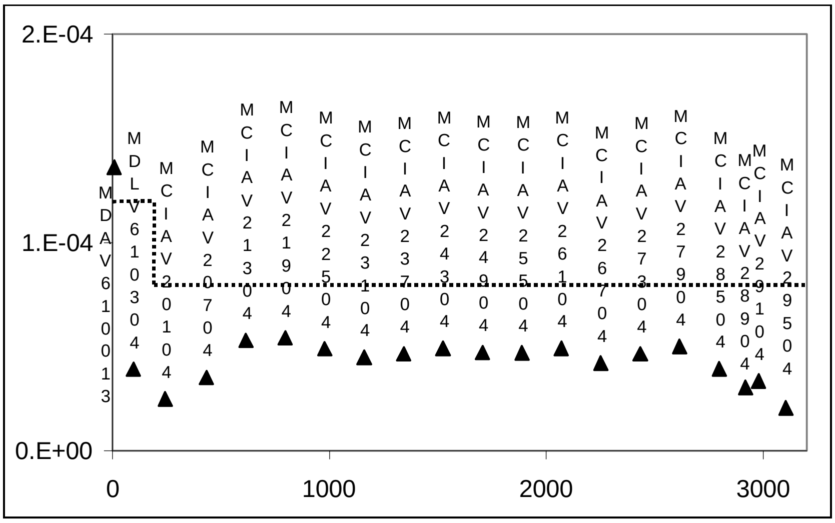


Figure 10: Triangles: corrector strengths in [rad] when the global error magnitude corresponds to 3 σ, assuming unlimited monitoring power, along the beam line in [m]; dotted line: physical corrector limits.

If the secondary response matrix $\mathbf{M}^{\text{resp}}_{\mathbf{CMA}}$ of (2.31) is used instead of $\mathbf{M}^{\text{resp}}_{\mathbf{CME}}$, the

[6] The evaluation of the monitor and corrector configurations should nonetheless not be understood as independent processes, as the configuration of one will definitely impact the effectiveness of the other in all but a few cases. Therefore an iteration of the process discussed here can be necessary.

corresponding evaluation will be basically independent of the monitor configuration and reflect a more fundamental correction range for each of the correctors. An example of this is shown in Figure 10. Here the triangles represent the corrector strength needed in units of radian at each corrector when the global error magnitude corresponds to 3 σ. The horizontal line represents the physical corrector limits.

2.5.4.2 Uncorrectable Orbit

The fundamentally uncorrectable orbit at all monitors is defined by the secondary response matrix $\mathbf{M}_{\mathbf{CM}}^{\text{uncorr}}$ of (2.32). This is a function at each monitor of both the monitor and corrector configurations, as shown in Figure 10 the 3 σ error envelope. The same fundamentally uncorrectable orbit at <u>all elements</u> can be obtained through $\mathbf{M}_{\mathbf{CMA}}^{\text{resid}}$ of (2.34). An example is shown in Figure 12. Finally, by using $\mathbf{M}_{\mathbf{CA}}^{\text{uncorr}}$ of (2.32) instead we can obtain the fundamentally uncorrectable orbit at all elements, as a measure purely of the corrector configuration even independent of the monitor configuration. An example of this is given in Figure 13.

2.5.4.3 Monitor Offset Induced Orbit Error

The secondary response matrix $\mathbf{M}_{\mathbf{MCA}}^{\text{ind}}$ defined in (2.33)

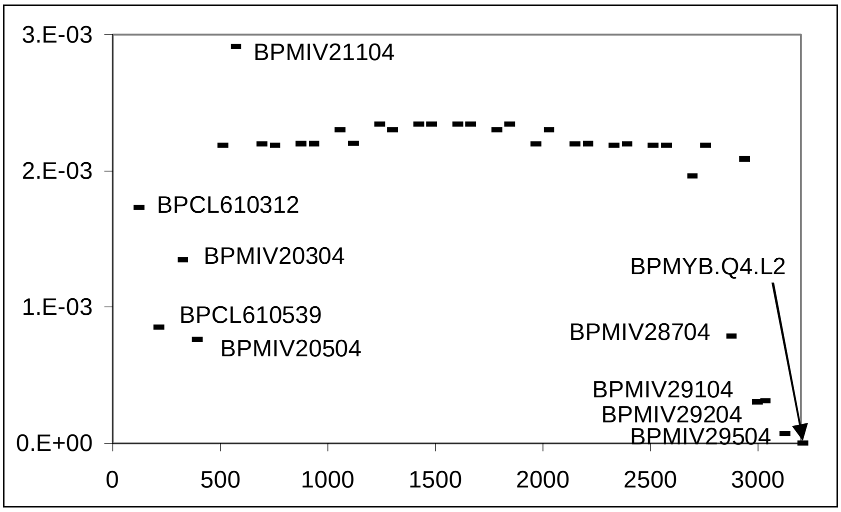


Figure 11: Uncorrectable orbit at all elements in [mm] along the beam line in [m].

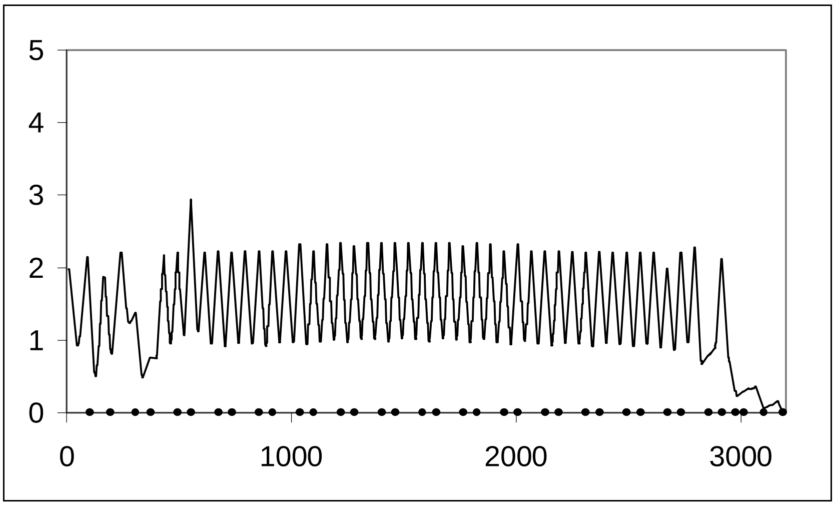


Figure 12: Full line: uncorrectable orbit at all elements in [mm], using existing monitor configuration along the beam line in [m]; dots: location of beam position monitors.

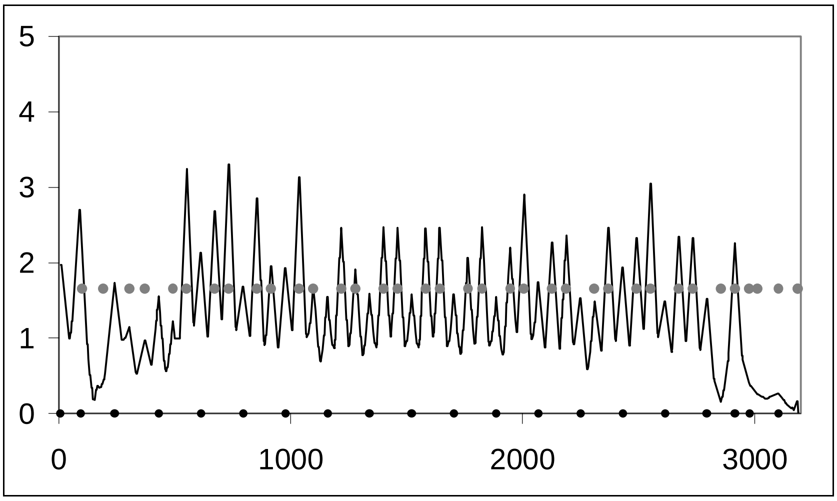


Figure 13: Full line: uncorrectable orbit at all elements in [mm], assuming unlimited monitoring power along the beam line in [m]; black dots: location of beam position monitors, grey dots: location of correctors

projects the normalised monitor error ellipsoid onto the element space, providing an evaluation of the effect of monitor offsets on the global orbit error. An example is given in Figure 14.

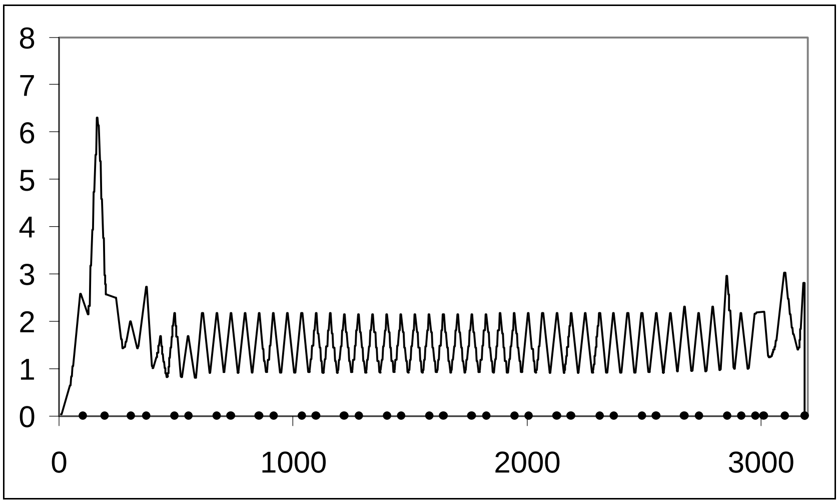


Figure 14: Full line: monitor offset induced orbit error in [mm] along the beam line in [m]; dots: location of beam position monitors.

2.5.4.4 Orbit Error Implied by Observed Orbit at Monitors

Using the secondary matrix $\mathbf{M}_{\mathbf{EC}}^{\text{rms-orb}}$ described in (2.36), we can evaluate the implication of an observed orbit, in terms of its RMS, for the underlying orbit error at all elements. The normalised monitor error ellipsoid is projected through $\mathbf{M}_{\mathbf{EC}}^{\text{rms-orb}}$ onto the space of all elements, which is demonstrated in Figure 15. This result is obtained after scaling down the envelope such that it corresponds to the original alignment type errors at a distribution not exceeding 3 σ at each point, using the cutoff schemes explained in section 2.3.5 and Appendix I.

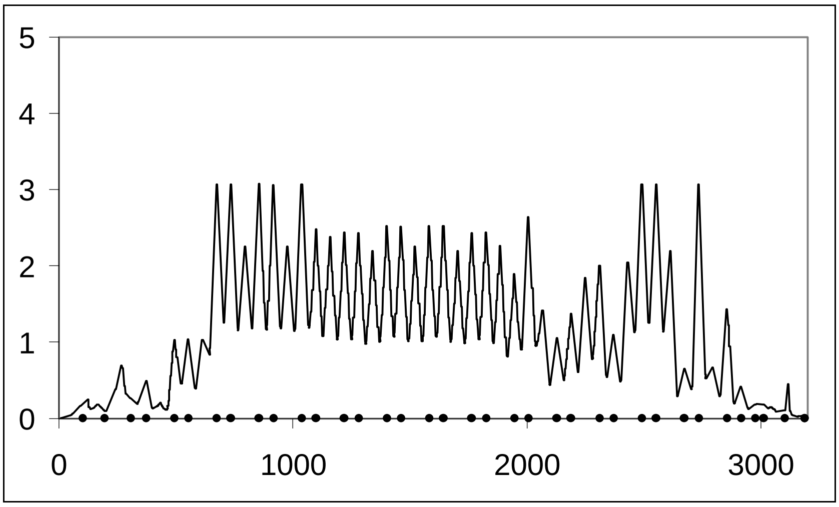


Figure 15: Full line: orbit error implied by observed orbit at the monitors, scaled by 3σ error, in [mm] along the beam line in [m]; dots: location of beam position monitors.

We can again take the quadratic sum of the two envelopes of Figure 12 and Figure 15, as they represent orthogonal components of the overall orbit error at compatible definition of probability density values. This is shown in Figure 16.

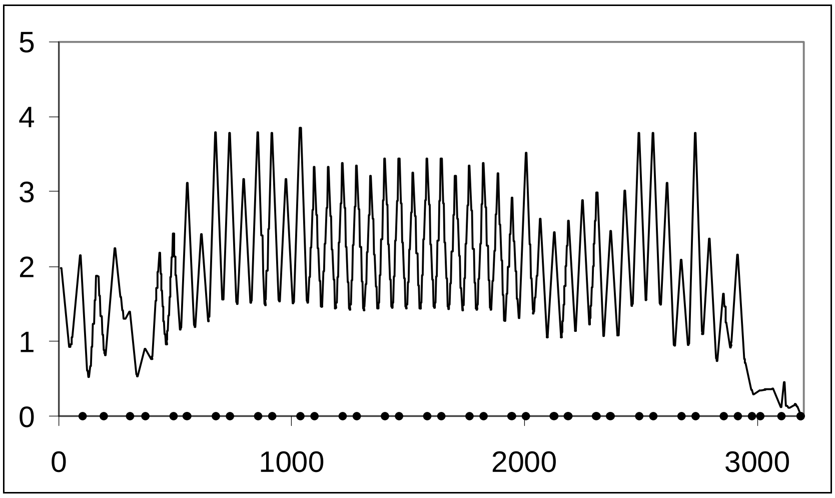


Figure 16: Full line: quadratic sum of the two previous orbit envelopes in [mm] along the beam line in [m]; dots: location of beam position monitors.

2.5.5 Actual Underlying Orbit Error

Using the secondary response matrix $\mathbf{M}_{\mathbf{EMA}}^{\text{UOE}}$ defined in (2.40), we can obtain the projection of an error ellipsoid, containing injection, alignment, field and monitor errors combined in a global probability distribution, onto each axe of the actual underlying orbit space $\mathbf{R}^{UA}$ defined in (2.40). This information alone summarises many important aspects of the orbit correction configuration, and can be meaningfully linked to a simulation test. An example is given in Figure 17 for the 3 σ envelope at all elements, where the 3 σ value for the exit angle distribution is also printed.

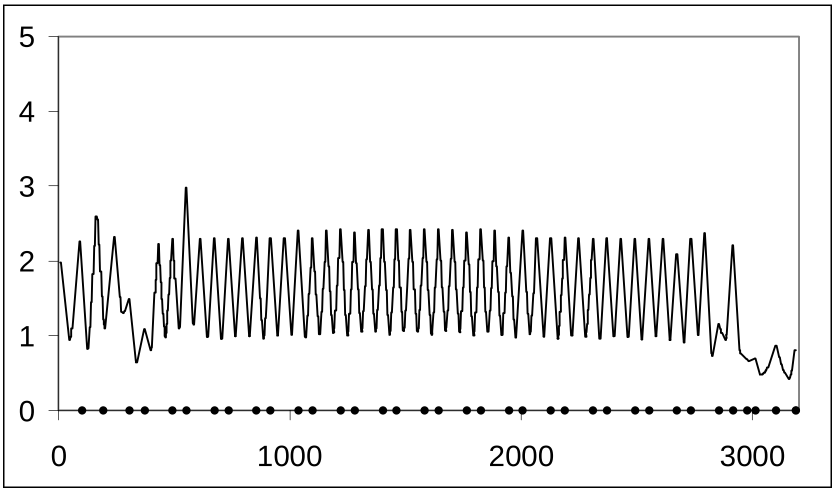


Figure 17: Full line: actual underlying orbit error in [mm] along the beam line in [m]; dots: location of beam position monitors.

2.5.6 Near Degeneracy in Response Matrices

By performing an SVD as in (2.24) on the response matrix $\mathbf{M}^{CM}$, we can evaluate the degree of degeneracy in a particular monitor and corrector configuration. The row vectors of $\mathbf{V}$ corresponding to singular values smaller than the cutoff defined by (2.27) represent near degenerate corrector combinations, as shown in Figure 18. Such combinations are dangerous in that they can lead to excessive correction strength and orbit error in unobservable locations. Completely degenerate combinations, corresponding to a zero singular value, are less likely in critically and over-constrained cases, but must be examined using the null-space method and eliminated before SVD is applied. A well-defined procedure [5] can be applied to eliminate degeneracy in a configuration iteratively.

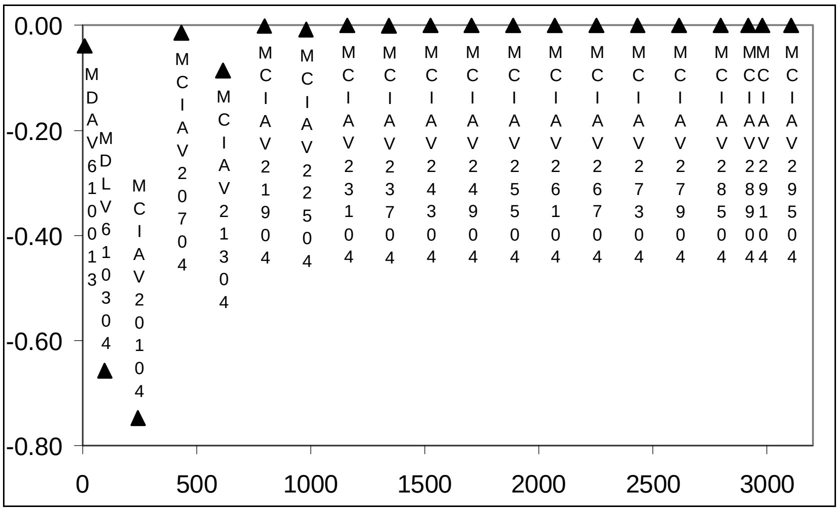


Figure 18: Example of near degenerate corrector with total magnitude normalised to 1 along the beam line in [m].

2.6 Study of Failure Modes

For most of the studies described in the previous section, whether on the 3 σ maximum along an axis or on the point of inscription between an ellipsoid and a hyper-cube, it is important to find not only the extreme or intersecting value, but also the error combination which led to such values. As explained in Appendices F, G and I, such configurations are obtained mostly through solutions to eigenvalue problems or inspection of mapping for Gaussian distributions. Examination of such error combinations provides useful insight leading to improved configurations. Figure 19 illustrates an effort to understand the combination of injection, alignment and monitor errors, as well as corresponding corrector responses that led to the calculated peak value for the actual underlying orbit. It is easy to see how various errors, including monitor error, conspire to create the extreme orbit excursion (clipped at graph

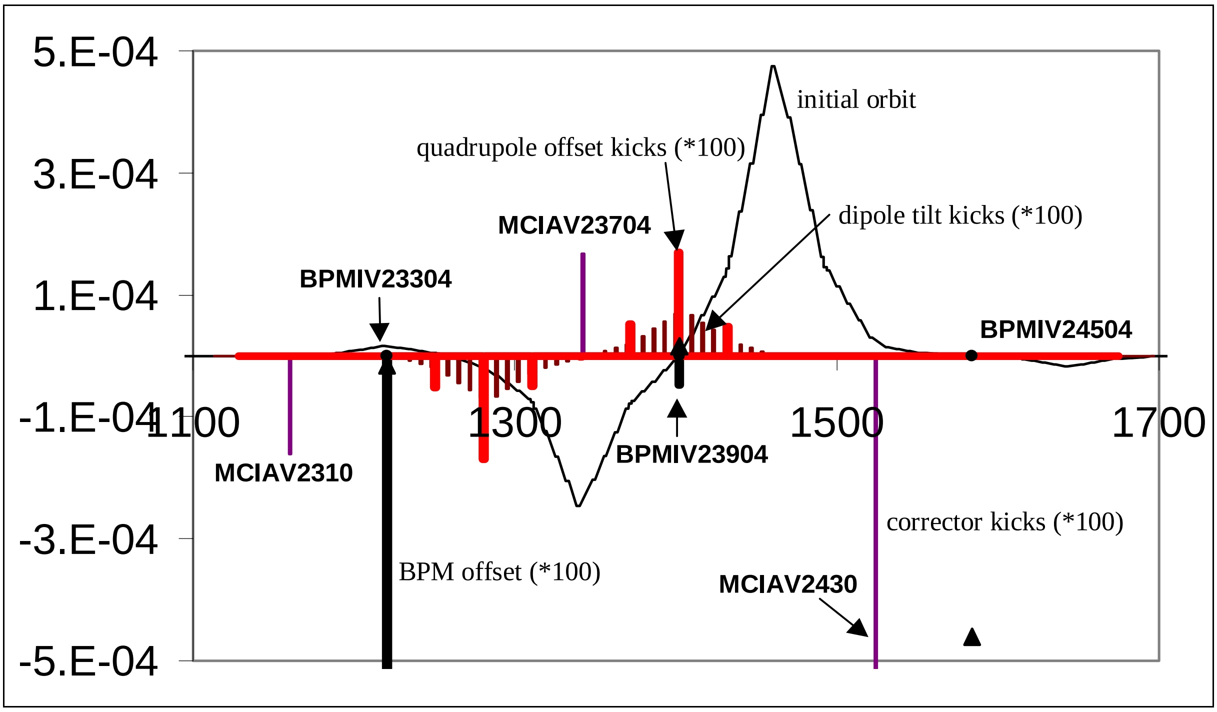


Figure 19: Example to study the composition of the leading underlying orbit (abscissa: beam line in [m]). (BPM offset and initial orbit in [m], all kicks in [rad])

boundaries). Such calculations are automatically generated by the analysis program and stored in files for later inspection.

The same studies are applied to corrector strength, observability and correctability limits. In each case the combination of errors leading to extreme conditions is calculated and examined as the need arises.

2.7 Configuration Optimisation

2.7.1 Improving the Configuration through Analytical Methods

Starting with a given configuration, we applied the analytical methods described in [5] to identify major problems in terms of observability, correctability and degeneracy. Iterative algorithms were applied to eliminate these problems to acceptable levels.

The iterative algorithms for eliminating degeneracy in monitor and corrector configurations have been described in sections 2.5.3 and 2.5.6. We describe below an algorithm for eliminating monitor deficiency, and two alternative algorithms for eliminating corrector deficiency. Armed with the ability to calculate uncorrectable residual orbit at all elements, the current approach is slightly different from that of [5].

2.7.1.1 Adding Monitors Based on the Unobservable Error-induced Orbit

First the error-to-monitor response matrix $\mathbf{M}^{\mathbf{EM}}$ is scaled so that in the actuator space all elements are measured in the σ's of the respective errors. Unobservability is indicated by the presence of null space or very small singular values of $\mathbf{M}^{\mathbf{EM}}$. In both cases the system fails to meet the observability criterion. The matrix $\mathbf{M}^{\mathbf{EM}}$ is decomposed with SVD. Those resulting error combinations corresponding to singular values short of a numerical criterion, defined through operational requirements, are identified. The error-to-all-element response matrix $\mathbf{M}^{\mathbf{EA}}$ is then taken to act on these error combinations to get error-induced orbits at all locations. The largest element of each of the orbit vectors is identified and, if this number exceeds a second numerical criterion for acceptable unobservable orbit, a new monitor is added at this location or its vicinity. The procedure is iterated until $\mathbf{M}^{\mathbf{EM}}$ no longer has singular error combinations whose effects in the all-element space exceed the second numerical criterion.

2.7.1.2 Adding Correctors Based on the Uncorrectable Residual Orbit

The maximal uncorrectable orbit at all locations is obtained through projection of the error ellipsoid by $\mathbf{M}_{\mathbf{CA}}^{\text{uncorr}}$ (defined in (2.32)). The eigenvalue method used for this purpose also gives the error combination $\mathbf{A}^{\mathbf{E,P}}$ leading to such maximum at each point P. Application of $\mathbf{M}_{\mathbf{CM}}^{\text{uncorr}}$ on the $\mathbf{A}^{\mathbf{E,P}}$ corresponding to the location with the largest not correctable orbit yields the residual orbit pattern under this particular error combination at all monitors

$$\mathbf{R}^{\mathbf{M,P}} = \mathbf{M}_{\mathbf{CM}}^{\text{uncorr}} \cdot \mathbf{A}^{\mathbf{E,P}}. \tag{2.44}$$

It remains to find the column vector in the generalised all-element-to-monitor response matrix $\mathbf{M}^{\mathbf{AM}}$ having the largest projection on $\mathbf{R}^{\mathbf{M,P}}$.

2.7.1.3 Adding Correctors Based on the Principal Axes of the Error Ellipsoid

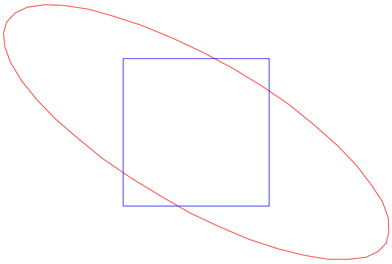


Figure 20A: Deficient corrector coverage of the monitor space.

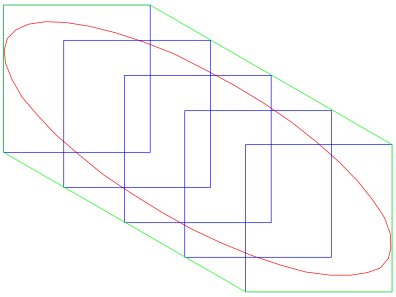


Figure 20B: Extended corrector coverage along the principal axis.

The above method works in cases over-constrained correction where $\mathbf{M}_{\mathbf{CM}}^{\text{uncorr}}$ is meaningful. There may be critically or even under-constrained cases in which we would like to add correctors to enhance correcting power. This situation is illustrated in Figure 20A, where the ellipse represents the projection of the 3 σ error ellipsoid onto the 2-dimensional monitor space. The rectangle represents the projection of the range of two correctors onto the monitor space. This is a critically constrained problem and there is no monitor subspace outside the column vectors of $\mathbf{M}^{\mathbf{CM}}$, thus $\mathbf{M}_{\mathbf{CM}}^{\text{uncorr}}$ does not exist. The system is fully correctable in the degree-of-freedom sense, but the corrector range is apparently deficient.

In this case we search for additional correctors with the purpose of enhancing the dimensionality of the hyper-cube representing the corrector range, such that when it is collapsed onto the monitor space (rectangular contour in Figure 20B) the added dimensions have the largest component along the principal axes of the projected error ellipse. The principal axes can be easily calculated by SVD as described in section 2.2.5. It remains to find the column vector in the generalised all-element-to-monitor response matrix $\mathbf{M}^{\mathbf{AM}}$ having the largest projection on the longest principal axis.

2.7.2 Exhaustive Fine Tuning of the Configuration

Analytical methods discussed thus far for improving the configuration are most effective for fixing major defects. When the question becomes fine tuning on trade-offs between a better beta function and a better phase advance when a corrector is moved by less than a few meters, the interplay between optics and error propagation becomes intractable analytically. In such cases it is best to use an automated process that scans over all possible monitor-corrector configurations[7], taking one of the quantities discussed in section 2.5[8] as a merit function. The configuration optimised by such a scan includes all the intractable interplay between observability, correctability and degeneracy based on the optics. It should be noted that such scans are only practical with an analytical calculation, as opposed to simulation, as the former can return a much faster and unambiguous result at each scanned configuration. It has been extensively used in fine tuning the currently proposed configurations and has proved very effective.

2.8 Identifying Critical Elements

Studies were made to measure the criticality of the elements in all proposed configurations. This is done by systematically deleting or altering each element in turn and examining the impact on the actual underlying orbit error. Since only analytical calculations are involved, scanning over all elements using this method is quite efficient and is routinely included in the standard analysis of a configuration. A total of four tests are done.

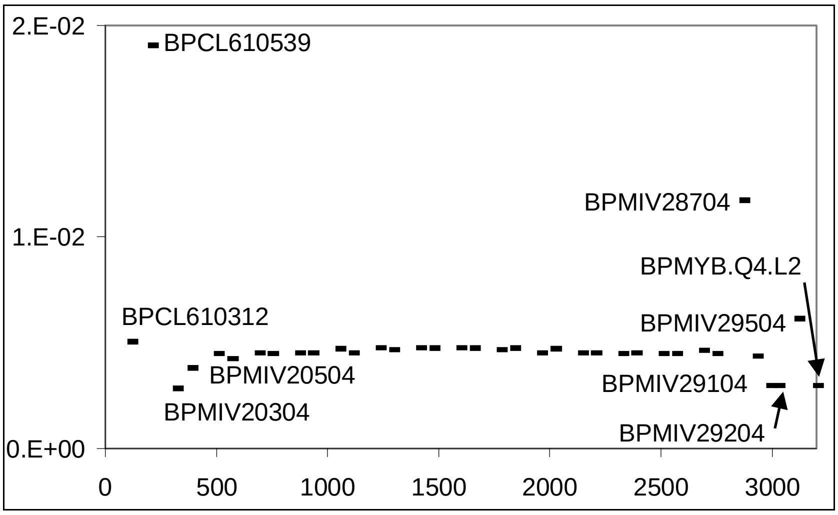


Figure 21: Maximum underlying orbit caused by missing monitor in [m] along the beam line in [m].

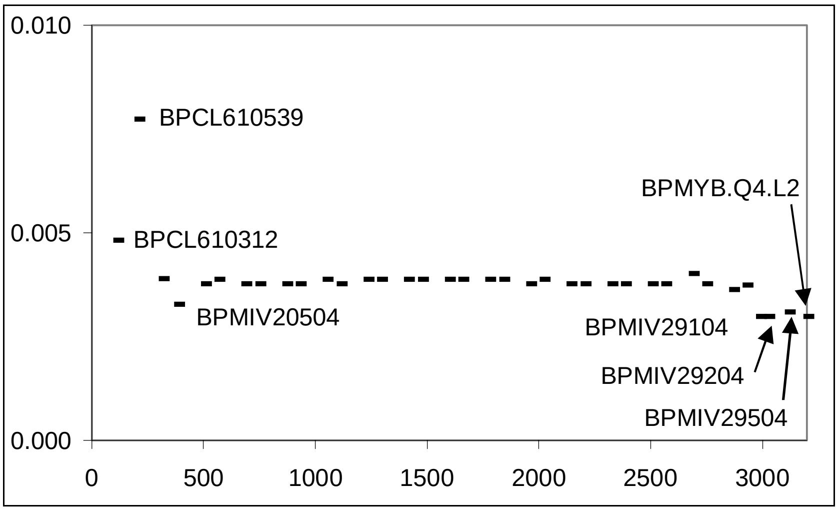


Figure 22: Maximum underlying orbit caused by fixed offset in [m] along the beam line in [m].

2.8.1 Impact of Missing Monitor

To examine the impact on the global envelope of the actual underlying orbit error due to a missing monitor, we use the secondary response matrix $\mathbf{M}^{\mathrm{UOE}}_{\mathrm{EMA}}$ as defined in (2.40) with the modification on $\mathbf{M}^{\mathbf{EM}}$, $\mathbf{M}^{\mathbf{CM}}$, $\mathbf{Z}^{\mathbf{MM}}$ and $\mathbf{M}^{\mathbf{MM}}$ having one of the monitor entries deleted. The resulting global peak in the actual underlying orbit error is plotted in Figure 21 as a function of the deleted monitor. It can be seen that the global peak remains the same when one of the monitors is missing from the entire set, except for one monitor towards the end whose absence would cause the global peak to increase. The program also calculates the exact global envelope in these cases to pinpoint the source of the problem, as shown in Figure 25. This is also done for the other criticality tests discussed below.

[7] Usually we scan over only one monitor and one corrector simultaneously over a limited range in the line. More elaborate scans would certainly be more time consuming.

[8] The ideal for such a merit function is the global actual underlying orbit.

2.8.2 Impact of Fixed Monitor Offset

To examine the impact on the global envelope of the actual underlying orbit error due to a fixed offset Δ at one monitor, we use the secondary response matrix $\mathbf{M}_{\mathbf{EMA}}^{\mathrm{UOE}}$ as defined in (2.40). The fixed monitor offset introduces a shift in the global 3 σ envelope by a fixed amount given by

$$\mathbf{R}_{\mathbf{i}}^{\mathbf{AO}} = \mathbf{M}^{\mathbf{CA}} \cdot \mathbf{M}_{\mathbf{CM}}^{\dagger} \cdot \mathbf{A}_{\mathbf{i}}^{\mathbf{MO}},$$

$$\mathbf{A}_{\mathbf{j}}^{\mathbf{MO}} = \begin{cases} 0, \mathrm{j} \neq \mathrm{i}, \\ \Delta, \mathrm{j} = \mathrm{i}. \end{cases} \tag{2.45}$$

The resulting global orbit error $\mathbf{R}^{\mathbf{AO}}$, in absolute value, is then superposed to the standard underlying orbit error at 3 σ. This is done for each monitor and the global peak is plotted as a function of the monitor with the offset. This is illustrated in Figure 22 with Δ = 3 mm.

2.8.3 Impact of Missing Corrector

To examine the impact on the global envelope of actual underlying orbit error due to a missing corrector, we use the secondary response matrix $\mathbf{M}_{\mathbf{EMA}}^{\mathrm{UOE}}$ as defined in (2.40) with the modification on $\mathbf{M}^{\mathbf{CM}}$ and $\mathbf{M}^{\mathbf{CA}}$ having one of the corrector entries deleted. The resulting global peak in the actual underlying orbit error is plotted as a function of the deleted corrector. This is shown in Figure 23. The one corrector causing drastic increase in the global envelope is located at the beginning of the line, the absence of which significantly reduces the ability of the system to contain injection errors, as shown in Figure 26.

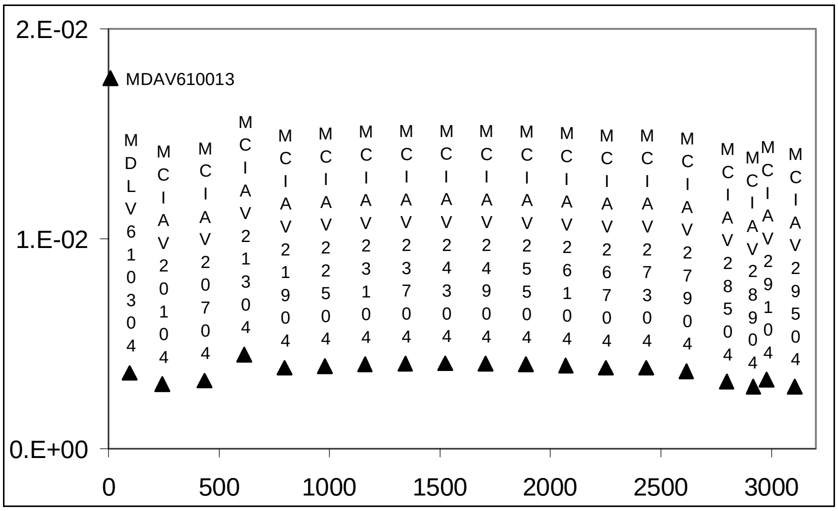


Figure 23: Maximum underlying orbit caused by missing corrector in [m] along the beam line in [m].

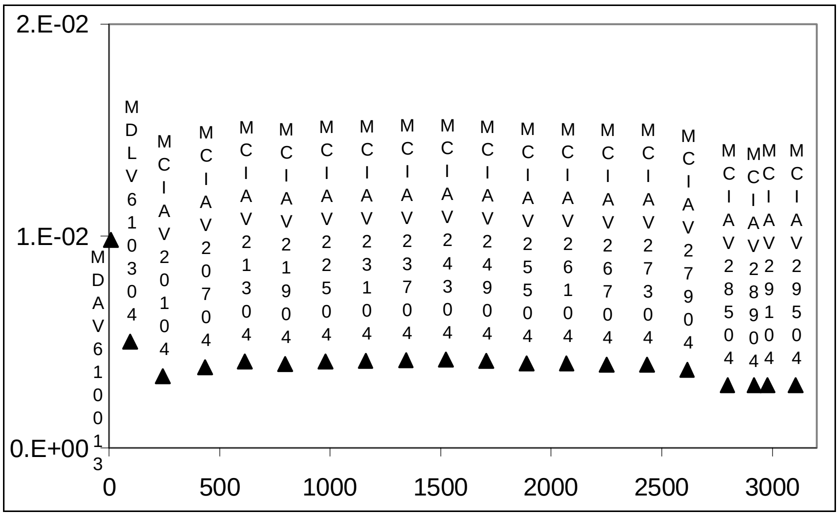


Figure 24: Maximum underlying orbit caused by fixed corrector scale error in [m] along the beam line in [m].

2.8.4 Impact of Fixed Corrector Scale Error

To examine the impact on the global envelope of actual underlying orbit error due to a fixed scale error Σ at one corrector, we use the secondary response matrix $\mathbf{M}_{\mathbf{EMA}}^{\mathrm{UOE}}$ as defined in

(2.40) with the modification on $\mathbf{M}^{\mathbf{CA}}$ having one of the columns magnified by (1+Σ). The resulting global peak in the actual underlying orbit error is plotted as a function of the anomalous corrector. This is shown in Figure 24 with Σ = 50 %.

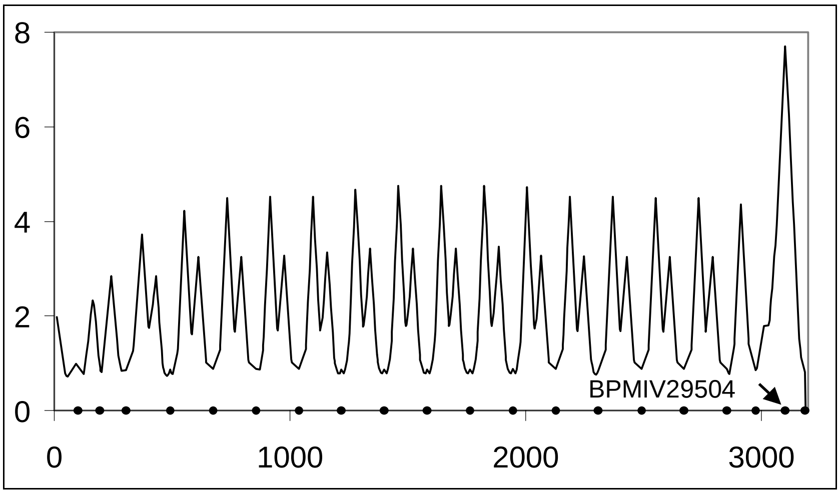


Figure 25: Leading underlying orbit in [mm] along the beam line in [m], with monitor BPIMV29504 missing.

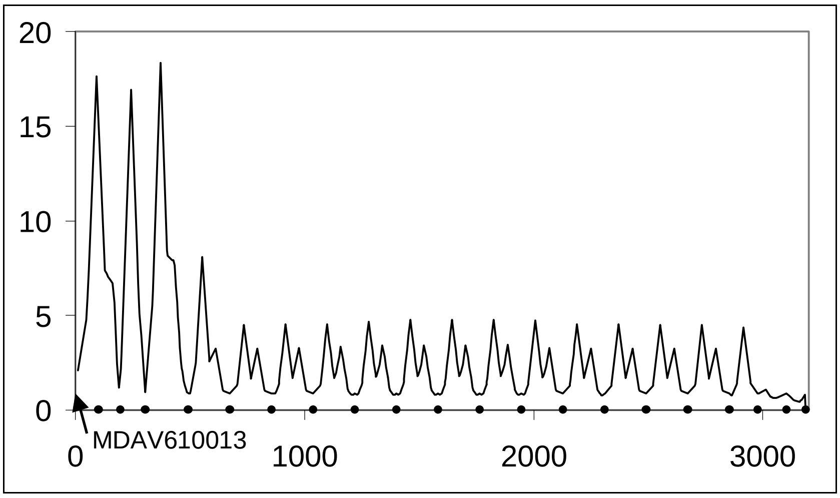


Figure 26: Leading underlying orbit in [mm] along the beam line in [m], with corrector MDAV610013 missing.

2.9 Comparison with Simulation

It is prudent, given the reliance of our analysis on analytical method, to perform some simulation as a cross-check. In the following results from massive simulations are given. In each example a total of 20,000 trajectories are launched, subject to the same optics and element configurations, error distributions, and orbit correction schemes as used by the analytical methods.

2.9.1 Envelope of the Actual Underlying Orbit Error

Figure 27 shows the 1 σ envelope of the actual underlying orbit error calculated by the analytical method (dots) and the simulation (full line). They demonstrate exact agreement at all elements.

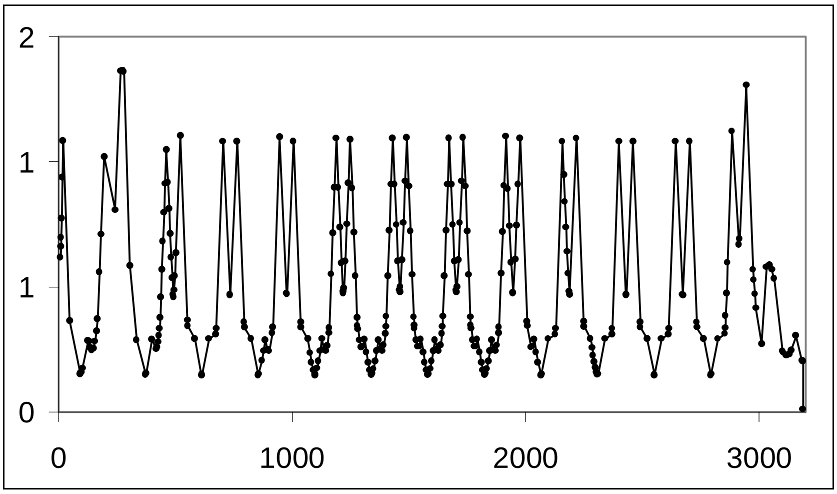


Figure 27: Maximum 1 σ underlying corrected orbit in [mm] along the beam line in [m] (dots: analytic method; full line: simulation).

2.9.2 Distribution of the Extremum

Figure 28 shows the distribution of the maximum actual underlying orbit error anywhere in the line, calculated by simulation. As explained in section 2.2.4 and Appendix I, no analytical method exists for evaluating this distribution. This however has the same characteristics as that displayed in Figure A.1, and agrees very well with an earlier simulation [2] of the extremum distribution in TI 2. Inspection of this and similar calculations carried out on various configurations seem to indicate that the extent of the extremum distribution is roughly 10 % - 20 % larger than the peak 3 σ envelope in the line, again in agreement with [2]. This rule-of-thumb can be used to roughly estimate the largest underlying orbit anywhere in the line, although there is no analytical justification for it.

2.9.3 The Distribution of the Error RMS and the Correlation with the Extrema of the Underlying Orbit Error

Figure 29 shows the distribution of the 20,000 maxima of the actual underlying orbit error anywhere in the line, correlated with their corresponding error magnitude measured in σ of the error distribution. As the total number of error sources is in the hundreds, we see the dramatic shift of the RMS peak predicted in Figure A.2 of Appendix I.

2.9.4 Direct Comparison of Orbit Error Distribution at Specific Points between Analytic Method and Previous Simulation

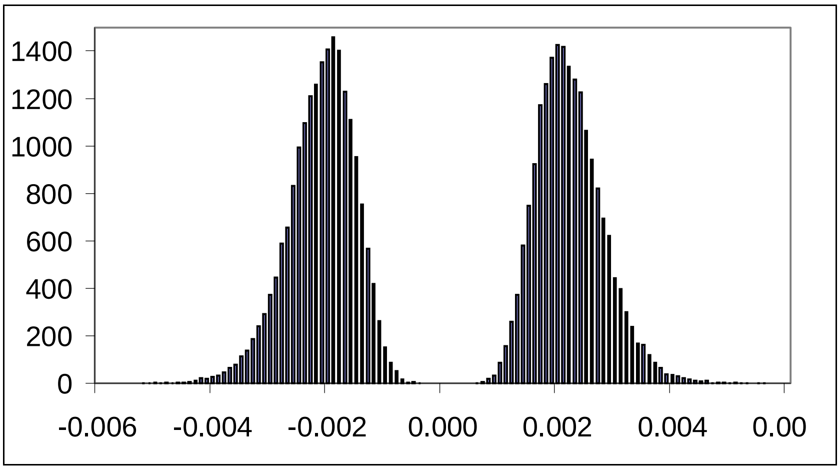


Figure 28: Distribution of maximum and minimum orbits in [m] for the corrected orbit.

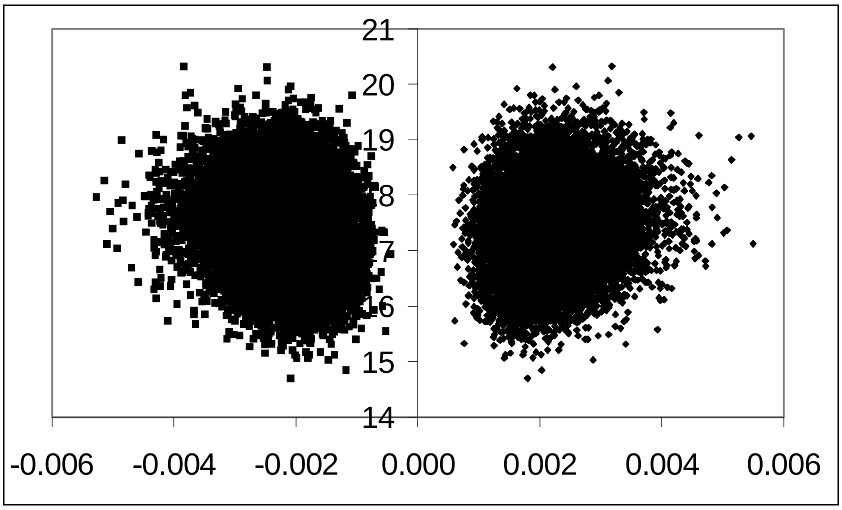


Figure 29: Correlation between corrected orbit maxima/minima and error sigma.

It should be noted that in the previous simulation study [2] the distribution used as the measure of performance was that of maximum orbit error values anywhere in the entire line for each random test, and thus follows a shape resembling that of Figure 28. Such a distribution, as explained earlier, cannot be represented in closed form analytically. A direct and unambiguous verification can nonetheless be made between the analytical results and the previous simulation, by comparing the 3 σ values of the orbit error distribution associated with specific points in the line. The 3 σ errors in both planes, at two representative points[9] in the TI 2 line one-in-three configuration, were calculated using both methods. The two methods agree on

[9] These are chosen at one focusing quadrupole and one defocusing quadrupole, resulting in four values very different in magnitude.

these four values to within 0.7% - 4.0%, consistent with the expected discrepancy due to statistics and known differences in inputs to these two methods[10].

[10] These include different injection errors, assignments of errors to dipoles and quadrupoles in non-periodic regions, models for monitor offset distributions, end angle constraints, position effects of dipoles used in steering, and difference in invoking one-to-one and SVD steering methods when steering units slightly overlap. The first three factors of the above are the more dominant ones and they all tend to make the analytical result slightly more pessimistic than simulation. This is indeed the tendency in the numerical differences seen in the comparison. The remaining factors are less dominant and their effects are more random than one-sided. Finally, a deviation by 4 % from the theoretical RMS in a sample of 1000 Gaussian-distributed values, equal in size to the sample used by the simulation, is quite normal.

3 The Result of the Analysis and the Proposed Configurations

3.1 Search for the Optimal Configuration and Evolution into the Over-Constrained Regime in the Periodic Sections

3.1.1 Procedure of Configuration Optimisation

The analytic methods described in section 2 were systematically applied to the lattices of the TI 2 and TI 8 lines in the following steps:

- A baseline configuration is constructed based on first-order considerations of optical properties such as betatron amplitude and phase advance, and reasonable balance between cost and performance.
- This configuration is then subjected to analysis by the program, which identifies defects in the configuration such as unobservability, uncorrectability, and near-singularities, in the meantime quantifying the various global performance parameters described in section 2,
- The configuration is improved, either by analytic methods when configuration defects are identified in the previous step or, in the case of configuration fine-tuning, through automated configuration search using global parameters as guidelines.

In the non-periodic parts of the two beam lines this procedure was followed through, with limited alternative, to arrive at the final configurations conforming to operational tolerances at minimal cost. In the periodic sections, it was clear that configuring monitors and correctors periodically in different ways could lead to interesting tradeoffs between performance and cost. A study was made to compare the merits of viable configurations in the periodic sections, the only requirement being their amenability to a 1-to-1 steering scheme. In other words, each corrector has to be identified with exactly one monitor in the steering process. The possibilities have also been studied of moving the periodic configuration boundary so that it extends further up or down-stream, and of shifting the relative pattern between monitors and correctors. These are summarised in Table 3.

Monitor pattern	**Corrector pattern**	**Average monitor per period**	**Average corrector per period**	**3σ orbit error peak (H/V mm)**	**Note**
2-in-4	2-in-4	0.50	0.50	3.2 / 3.4	
2-in-4	2-in-4	0.50	0.50	3.2 / 3.4	Monitors shifted by 1 optical cell
2-in-4	2-in-4	0.50	0.50	4.0 / 4.0	Configuration extended downstream
1-in-2	1-in-2	0.50	0.50	>10.0	Near singularity due to 90° cell
1-in-3	1-in-3	0.33	0.33	4.7 / 4.7	
1-in-4	1-in-4	0.25	0.25	8.0 / 8.0	

Table 3: Comparison between various 1-to-1 schemes in the periodic section.

It was concluded that the 2-in-4 scheme provides a sound balance between performance and economy insofar as 1-to-1 schemes are concerned. This is consistent with the conclusion of the previous study [2].

3.1.2 Evolution into the Over-Constrained Regime

In parallel with the evaluation of various 1-to-1 configurations, a study into over-constrained steering was initiated. This was motivated by the observation that, despite the relative advantage displayed by the 2-in-4 configuration, some near-singularity in the orbit correction was observed, indicating an overly dense coverage by correctors.

One such near-singular combination is shown in Figure 30 for TI 2. The same pattern repeats throughout the entire periodic section. Such near-singularities can result in excessive correction, both in terms of corrector strength and orbit error at locations not equipped with beam position monitors. These problems are not unusual, and are generally solved by reducing the corrector coverage and resorting to software-aided over-constrained steering.

The immediate path to corrector reduction in the 2-in-4 pattern is either a 1-in-3 or a 1-in-2 scheme. The latter would present even more severe singularity problems due to the 90°-per-cell optics.

Thus the most obvious option for corrector reduction was the 1-in-3 pattern, while maintaining the 2-in-4 pattern for monitors. Figure 31 shows the resulting 3 σ orbit error envelope for such a

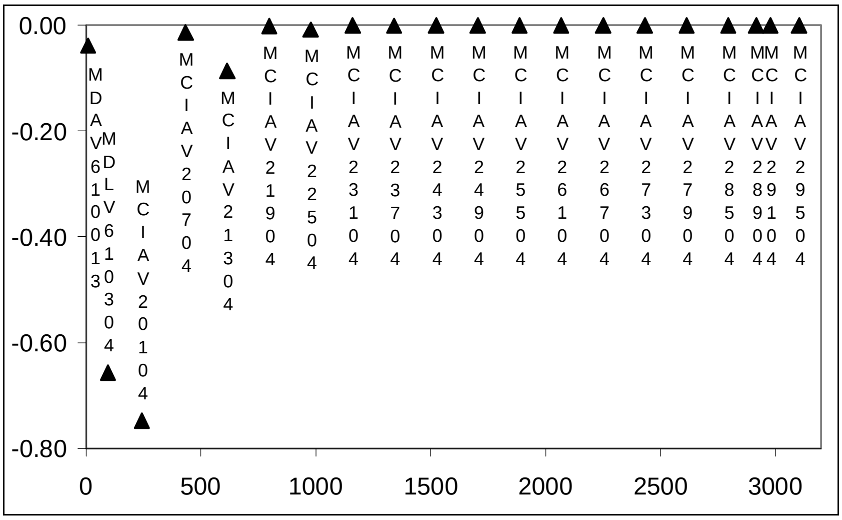


Figure 30: Triangles: corrector strengths in [rad] for a near-singular combination in the 2-in-4 scheme (TI 2; abscissa in [m]).

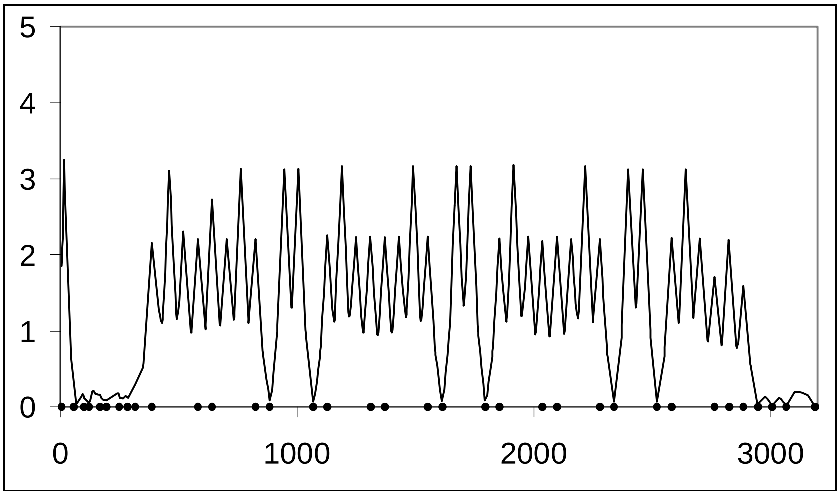


Figure 31: Full line: 3 σ orbit envelope in [mm] when applying an SVD-style correction on a 2-in-4 monitor and 1-in-3 corrector scheme (TI 2; abscissa in [m]); dots: position of monitors.

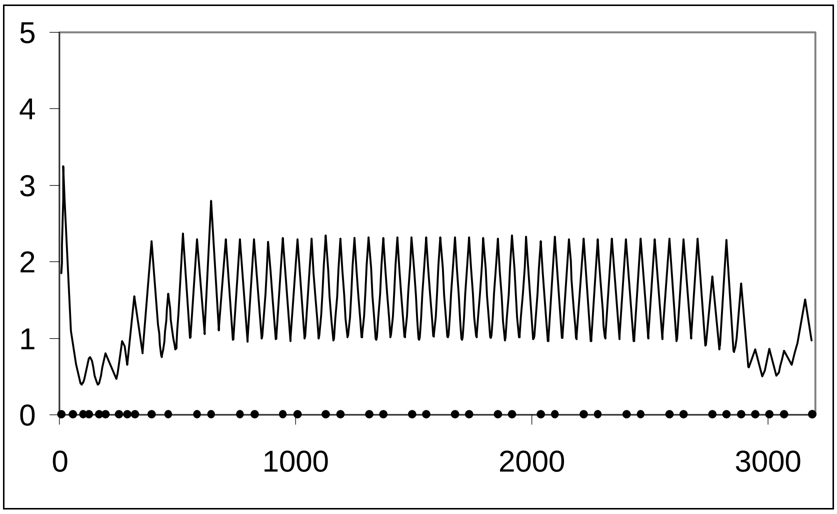


Figure 32: Full line: 3σ orbit envelope in [mm] for an over-constrained 2-in-3 monitor and 1-in-3 corrector scheme, (TI 2, abscissa in [m]; dots: position of monitors).

configuration when an SVD-style correction is applied to minimise the RMS orbit[11]. The corrector singularity has been reduced in this case, but inspection of Figure 34 reveals that, over each super-period of 12 cells, there is a clear sign of over-correction at some monitors at the expense of under-correction at other, mostly unmonitored locations.

The overall orbit error envelope can be improved if this imbalance is mitigated. This was achieved by shifting the monitors at the over-corrected locations to the gaps where the orbit is under-monitored and under-corrected, at the expense of two additional monitors per 12-cell super-period. The outcome turned out to be exactly a 2-in-3 pattern for the monitors. The orbit error envelope for this over-constrained configuration is shown in Figure 32, where the peaks in the periodic section have been reduced to 2.3 mm, a reduction of nearly 30%. More significantly, the entire periodic section is steered evenly, without unjustified emphasis and therefore waste of correction resources in particular local regions. The new over-constrained configuration, with a 2-in-3 scheme for the monitors and a 1-in-3 scheme for the correctors, is presented as alternative for both lines. Per 12-cell super-period, it requires two additional monitors but offers a reduction by two correctors. Anticipating the likely need to perform the initial orbit set-up during commissioning without computer support, an analysis has also been performed on a 1-to-1 variation of this scheme, with every other monitor disabled in the periodic section to form a 1-in-3 pattern in both monitors and correctors over the entire line.

3.2 Details of the Proposed Configurations

On the following pages two plots, separately per plane, are shown for each configuration optimised under the respective baseline theme for monitor and corrector placements. For both TI 2 and TI 8 the following configurations were optimised:

- 2-in-4 for both monitors and correctors;
- 2-in-3 for monitors and 1-in-3 for correctors;
- 1-in-3 for monitors and 1-in-3 for correctors.

For each configuration the first plot shows the individual corrector limit in units of the error σ that the corrector in question can handle. Triangles located below the dotted 3 σ line indicate those correctors whose limit will be reached before the error distribution reaches the surface corresponding to the 3 σ cutoff on all errors. It should be noted that the values above 3 σ are meaningful only if those below are “fixed” so that they no longer “max out” at 3 σ. However the plot does give an unambiguous ordering of the relative margins of the correctors, as explained in Appendix F.

The second plot shows the 3 σ extent at each representative element in the line of the true underlying orbit when the combined injection, alignment, and monitor offset errors reach the distribution surface of 3 σ. The dots along the horizontal axis mark the monitor positions. The 3σ envelope value of the outgoing angle is also given, as it is important in some cases.

[11] The effect of monitor offset errors is not included in this graph to accentuate the unbalanced orbit monitoring. Inclusion of monitor offset errors will make the global orbit error larger than displayed.

3.3 TI 2

3.3.1 TI 2: 2-in-4 for monitors, 2-in-4 for correctors, horizontal plane

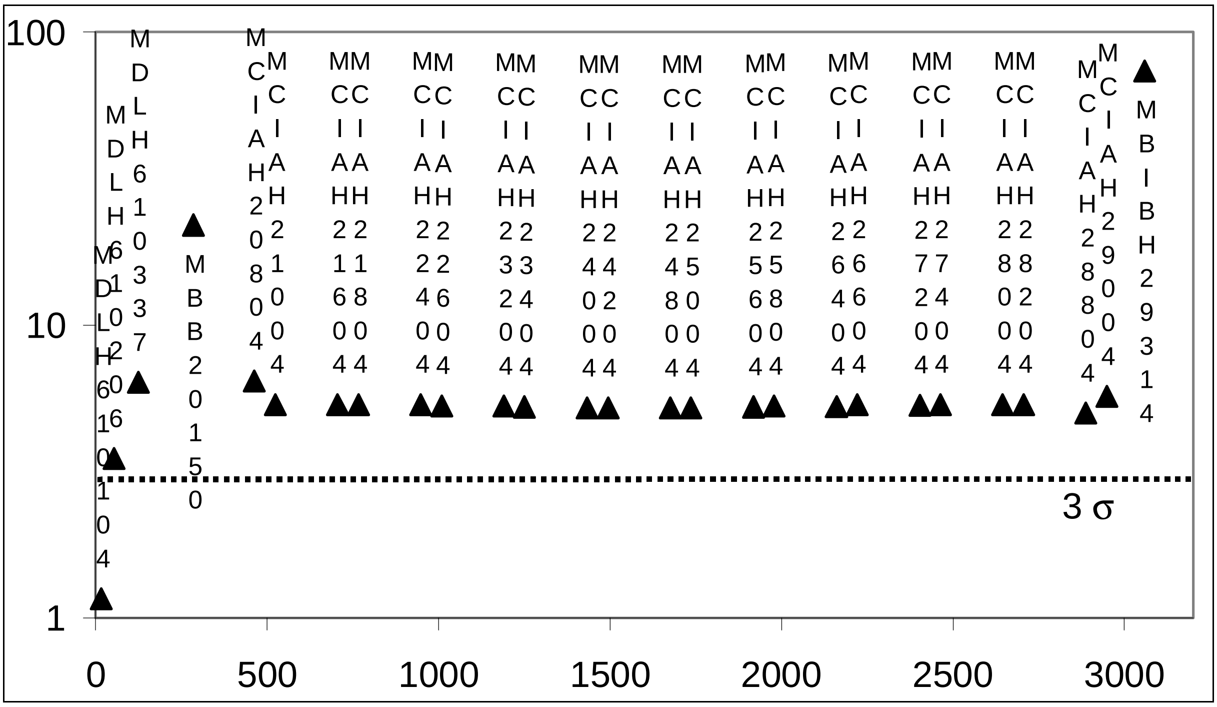


Figure 33: Corrector limits in error σ.

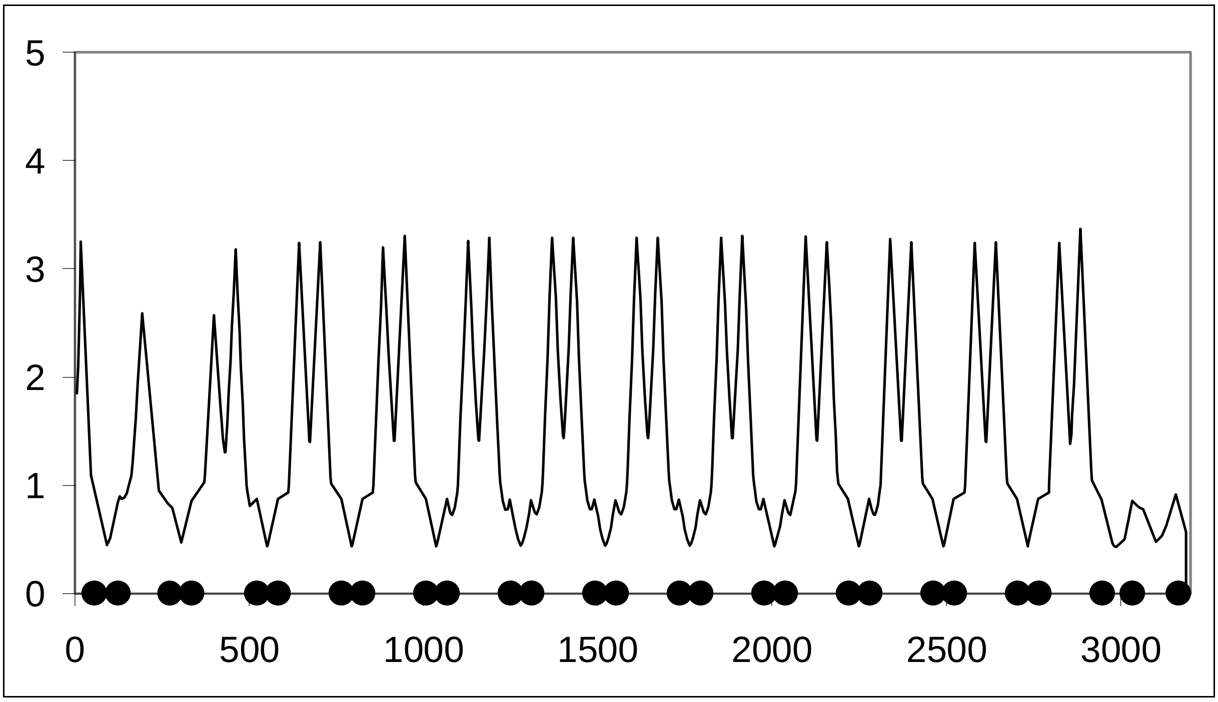


Figure 34: Resulting orbit error envelope in [mm] along TI 2 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 26 [μrad].

3.3.2 TI 2: 2-in-4 for monitors, 2-in-4 for correctors, vertical plane

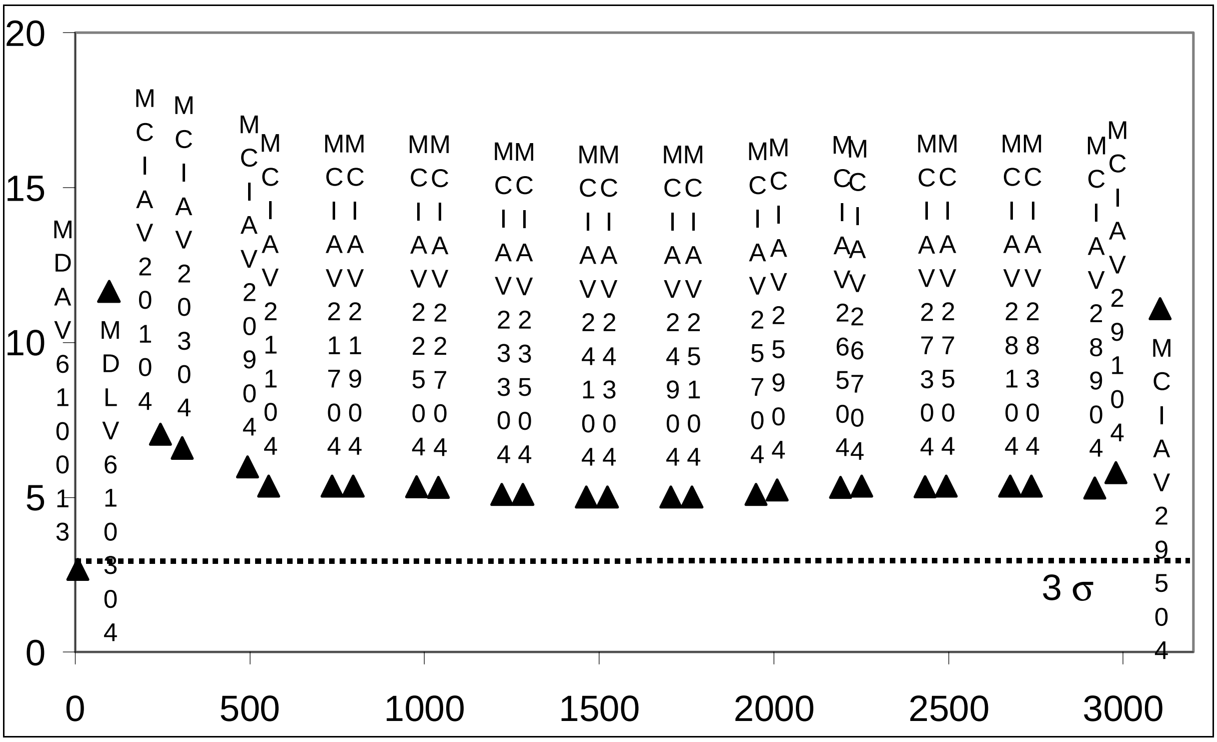


Figure 35: Corrector limits in error σ.

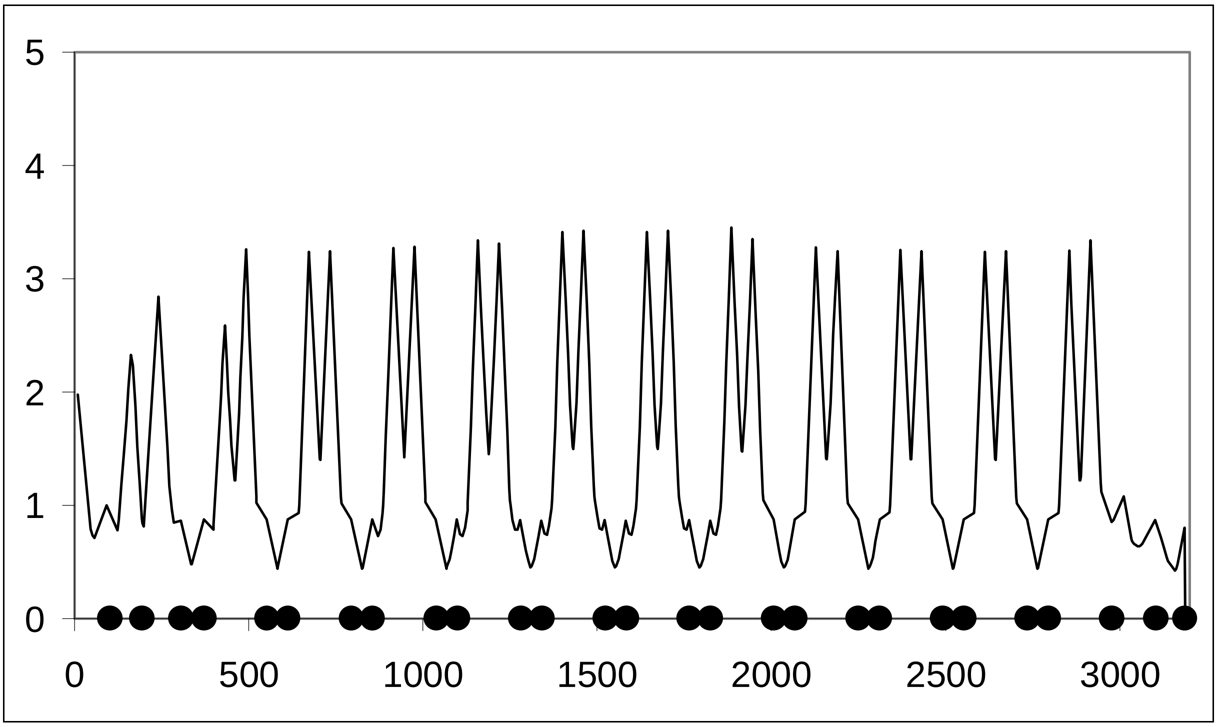


Figure 36: Resulting orbit error envelope in [mm] along TI 2 (in [m]; dots: monitor positions). 3 σ envelope value of the outgoing angle = 22 [μrad].

3.3.3 TI 2: 2-in-3 for monitors, 1-in-3 for correctors, horizontal plane

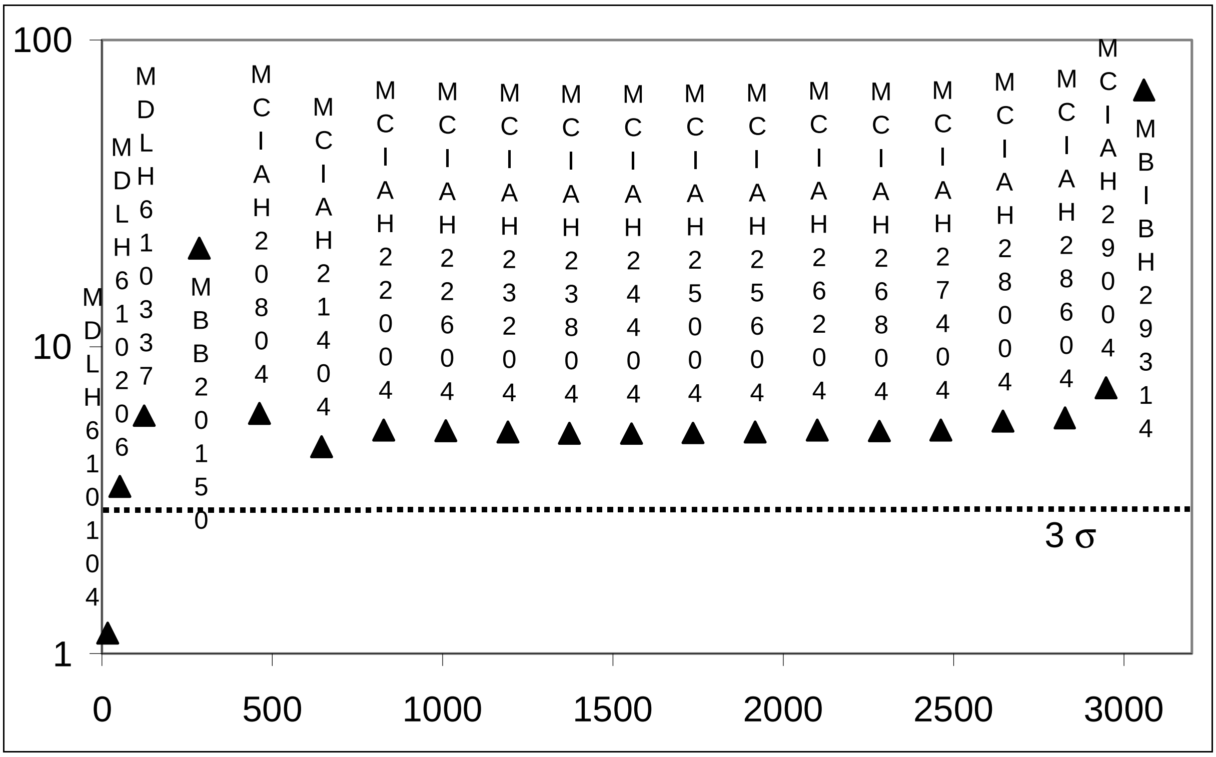


Figure 37: Corrector limits in error σ.

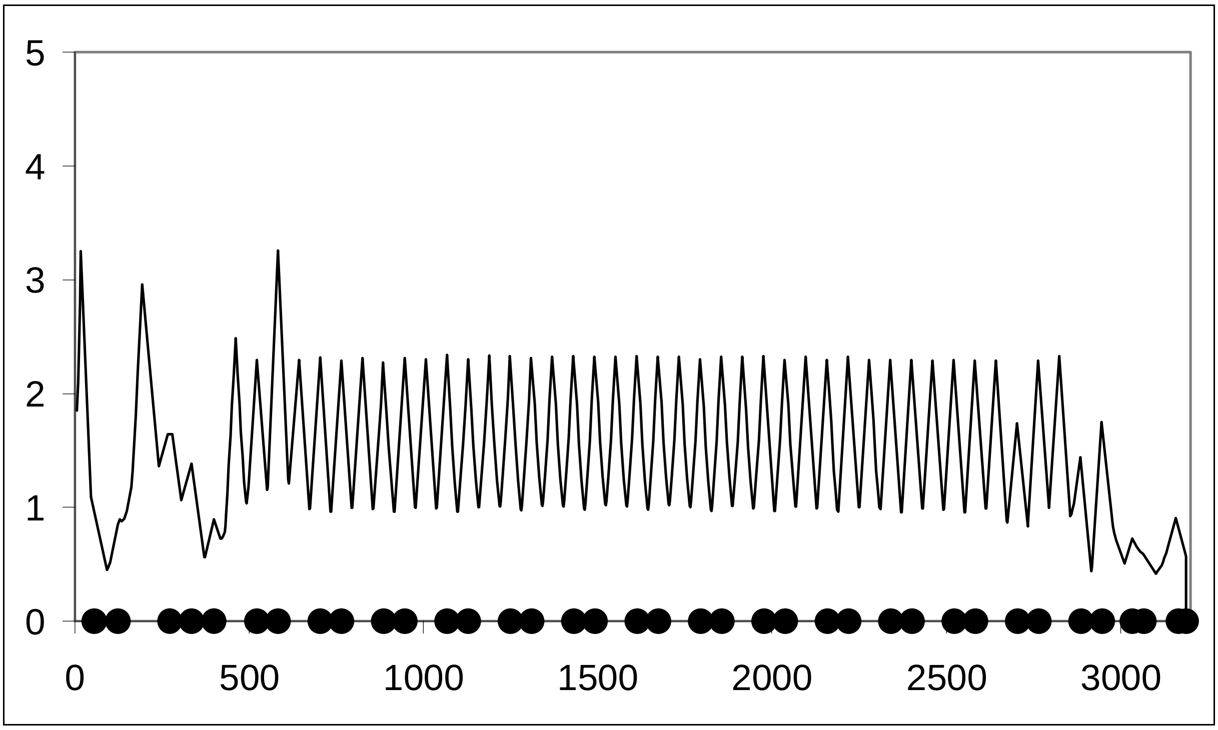


Figure 38: Resulting orbit error envelope in [mm] along TI 2 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 26 [μrad].

3.3.4 TI 2: 2-in-3 for monitors, 1-in-3 for correctors, vertical plane

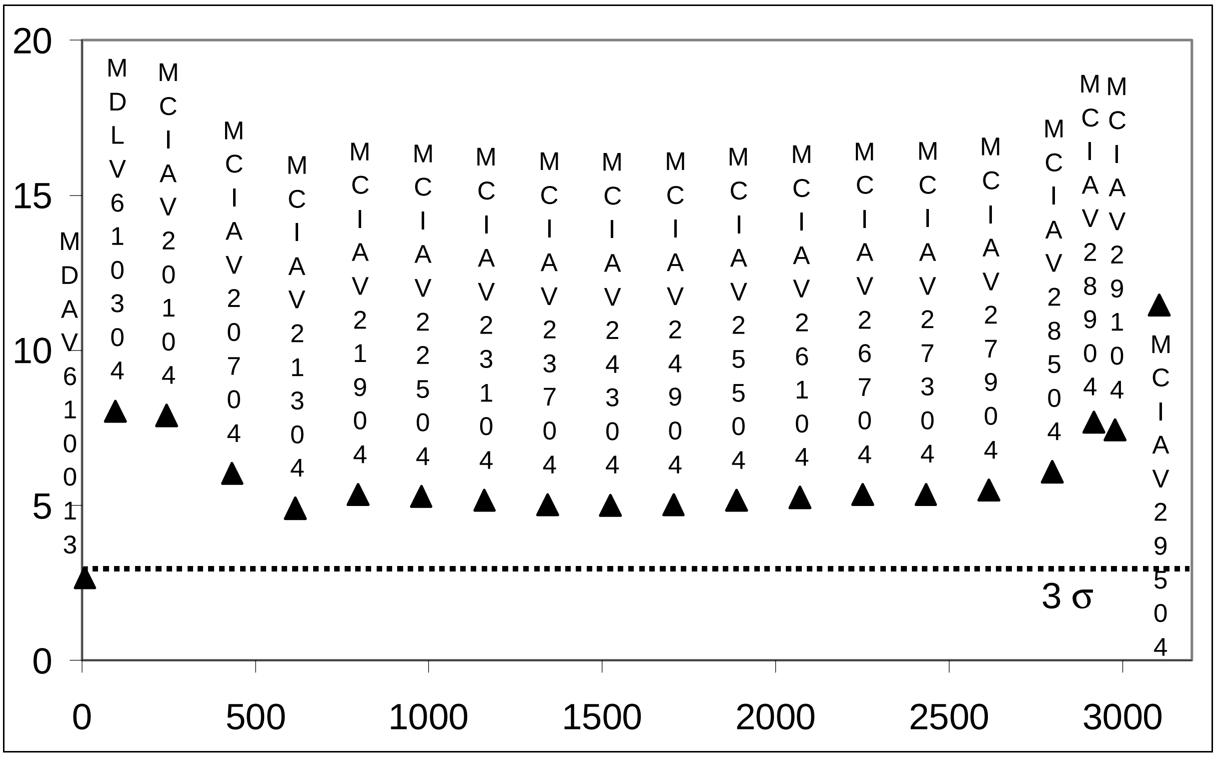


Figure 39: Corrector limits in error σ.

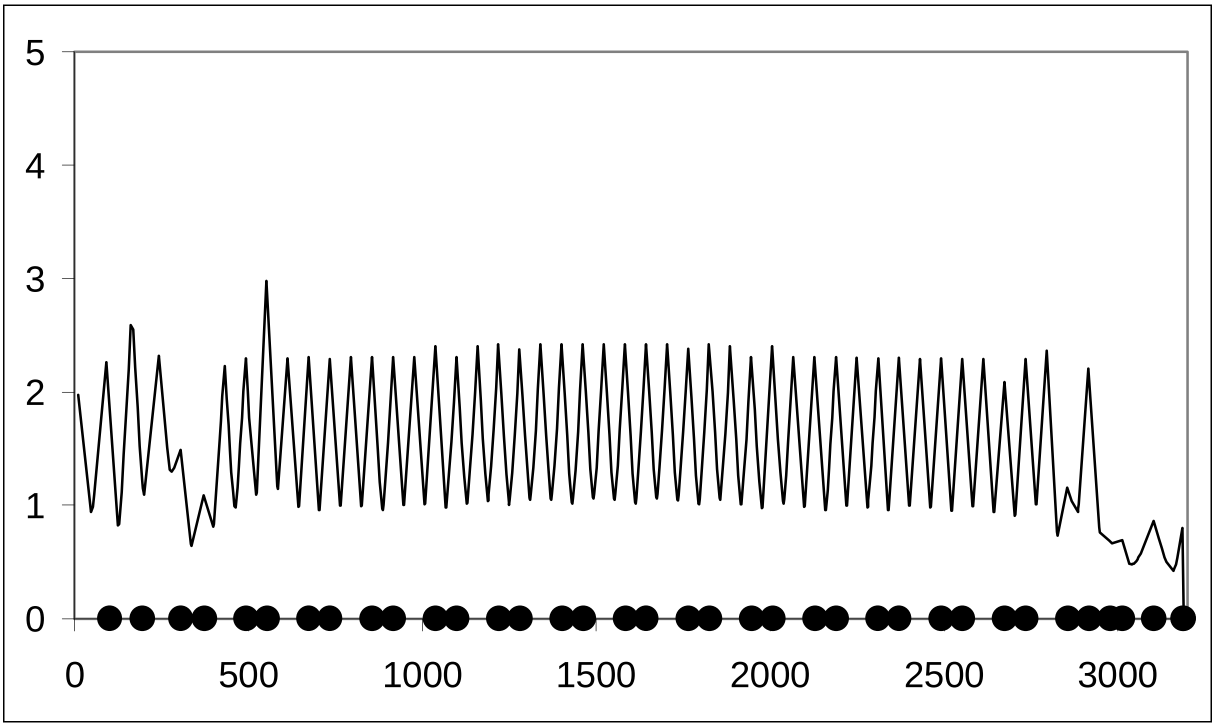


Figure 40: Resulting orbit error envelope in [mm] along TI 2 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 22 [μrad].

3.3.5 TI 2: 1-in-3 for monitors, 1-in-3 for correctors, horizontal plane

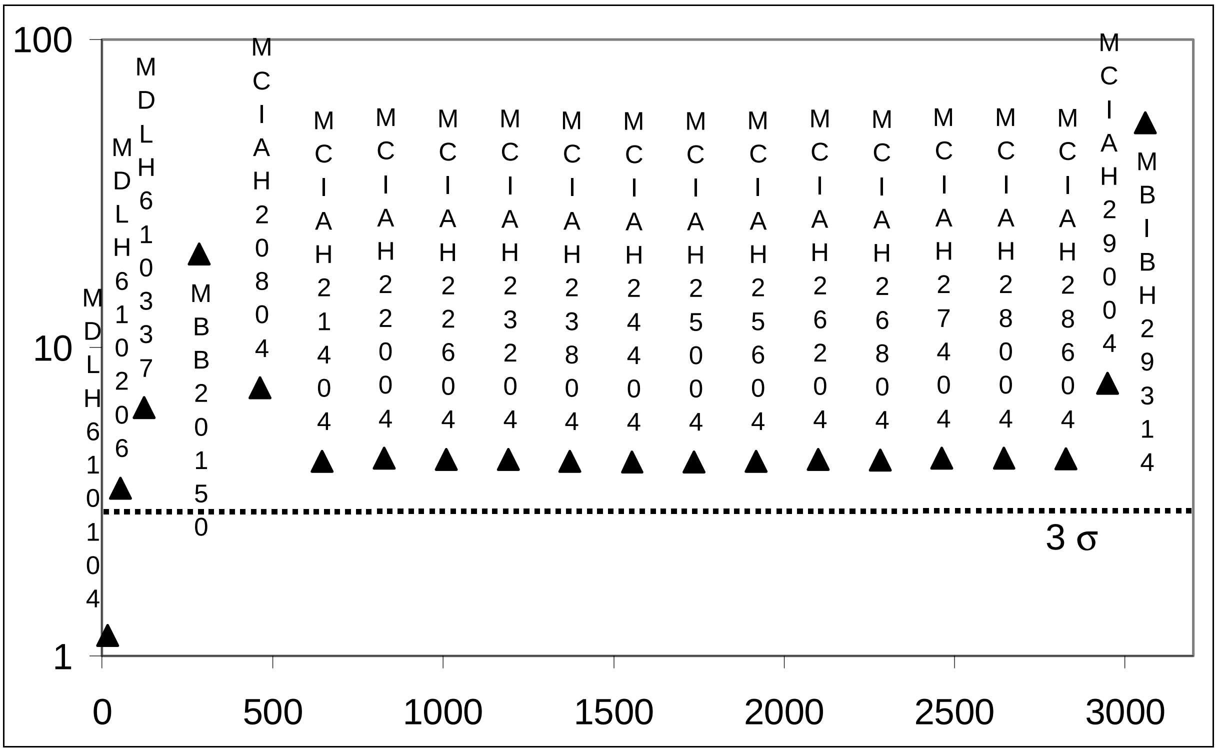


Figure 41: Corrector limits in error σ.

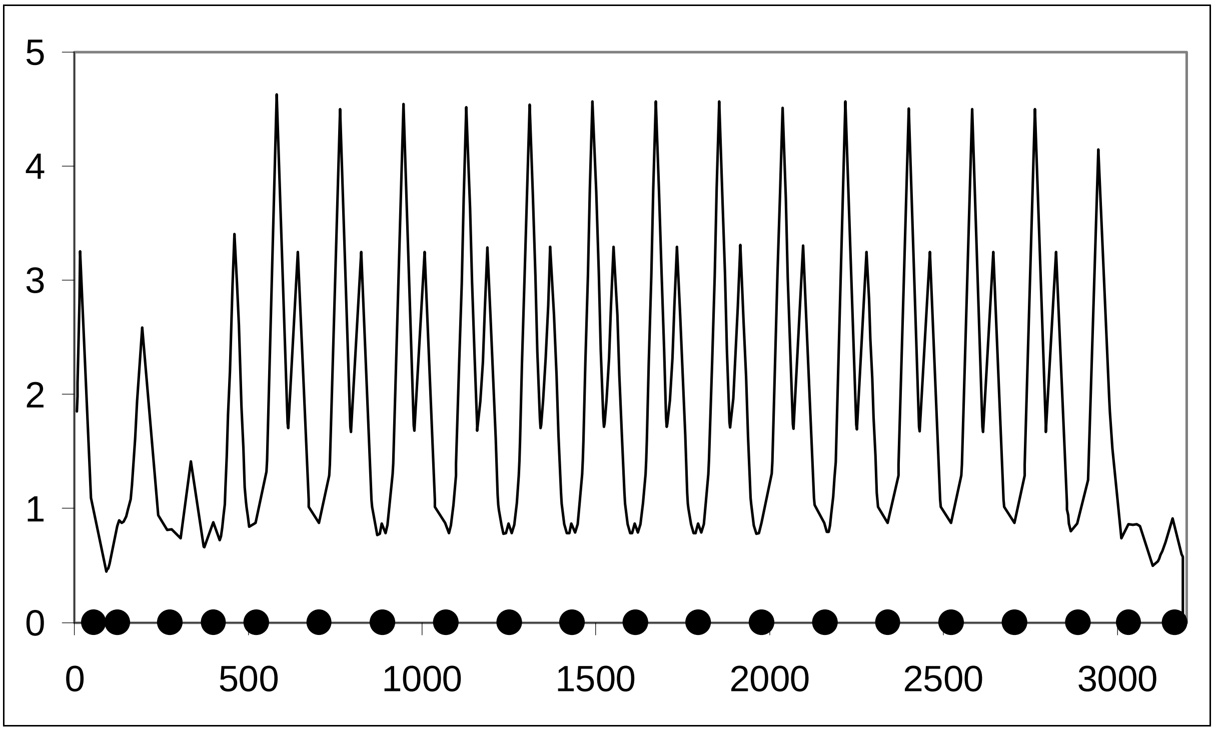


Figure 42: Resulting orbit error envelope in [mm] along TI 2 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 26 [μrad].

3.3.6 TI 2: 1-in-3 for monitors, 1-in-3 for correctors, vertical plane

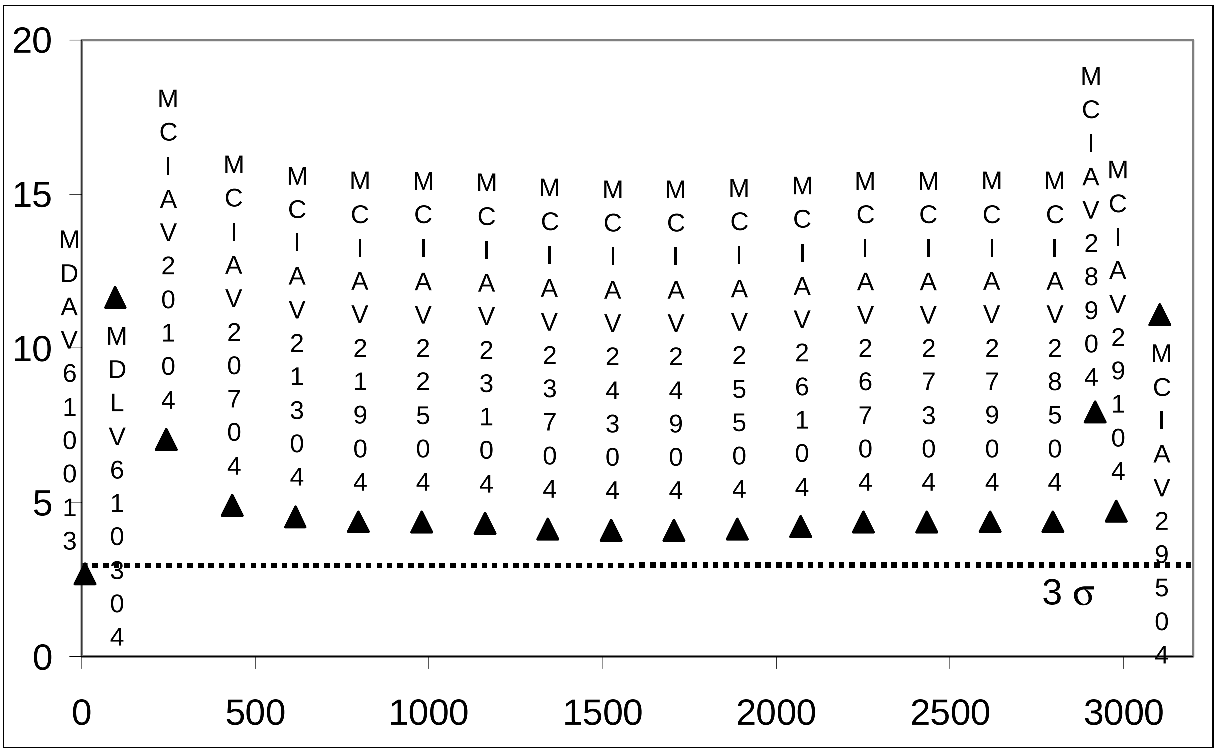


Figure 43: Corrector limits in error σ.

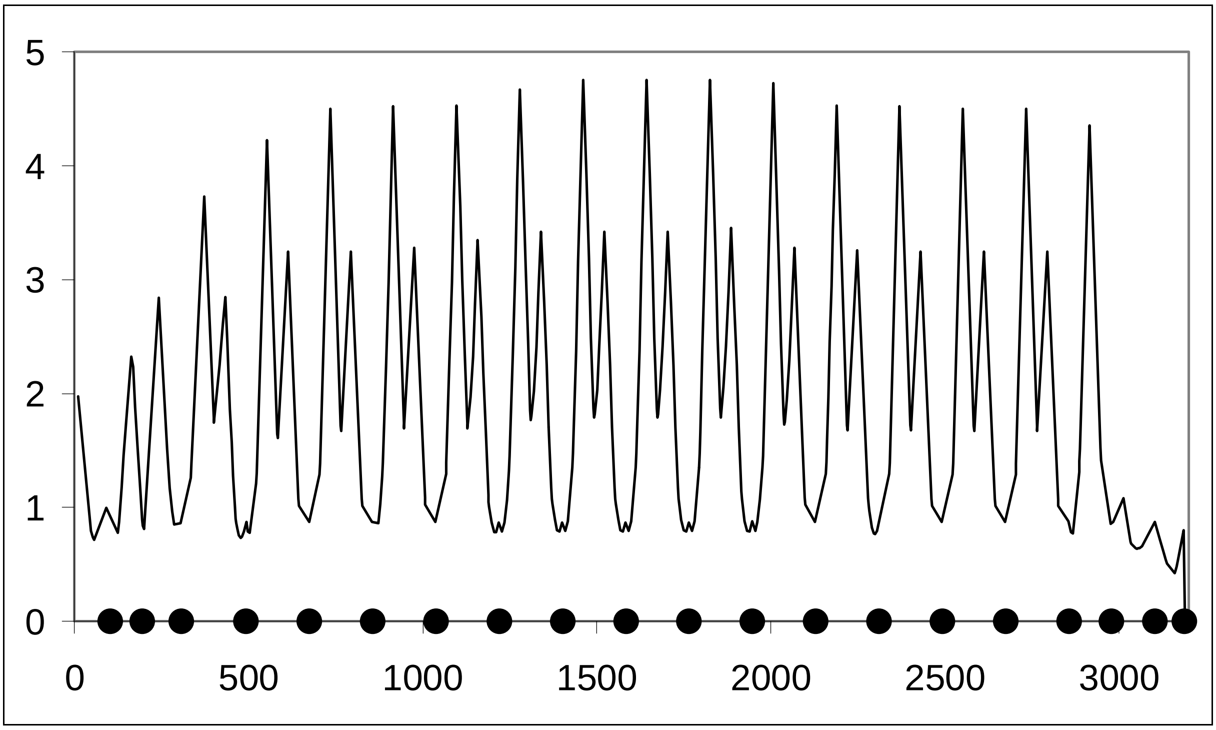


Figure 44: Resulting orbit error envelope in [mm] along TI 2 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 22 [μrad].

3.4 TI 8

3.4.1 TI 8: 2-in-4 for monitors, 2-in-4 for correctors, horizontal plane

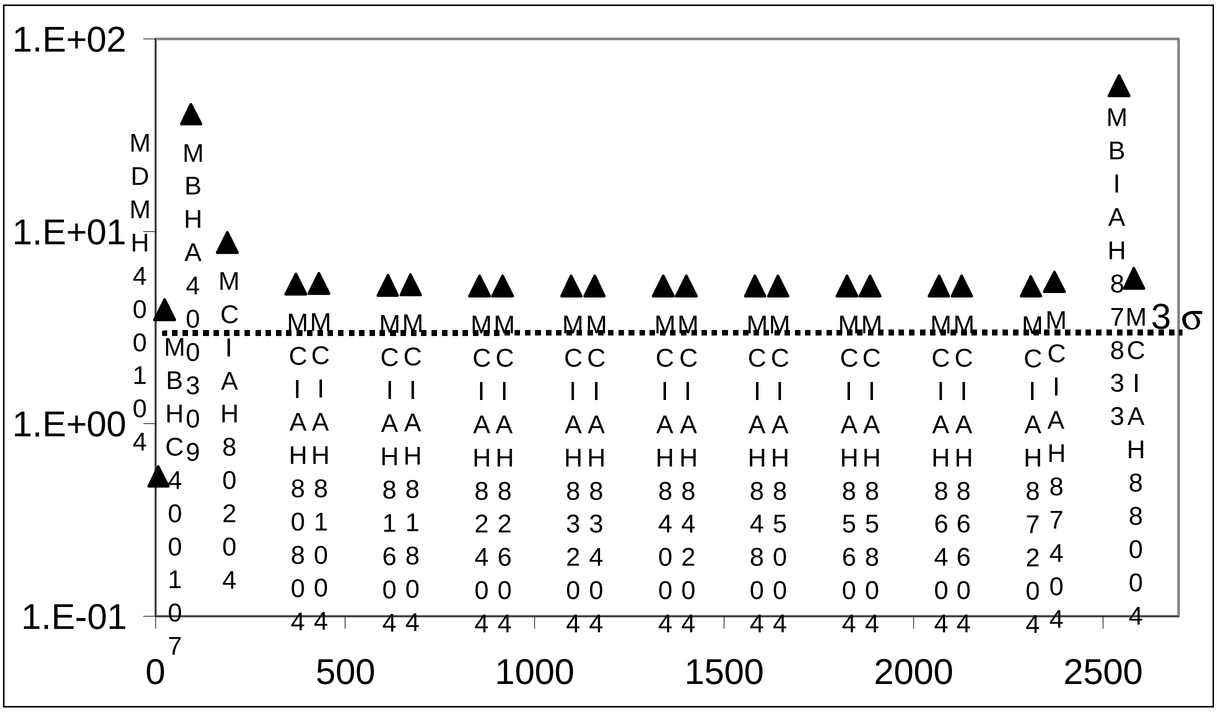


Figure 45: Corrector limits in error σ.

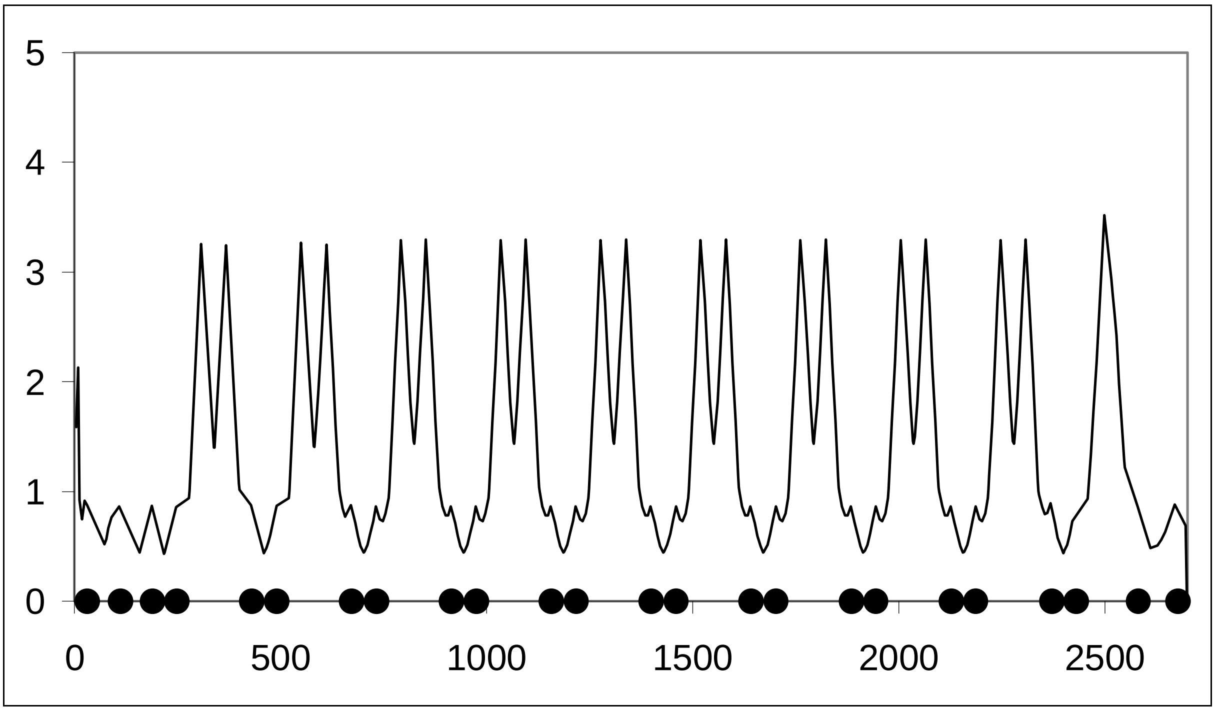


Figure 46: Resulting orbit error envelope in [mm] along TI 8 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 20 [μrad].

3.4.2 TI 8: 2-in-4 for monitors, 2-in-4 for correctors, vertical plane

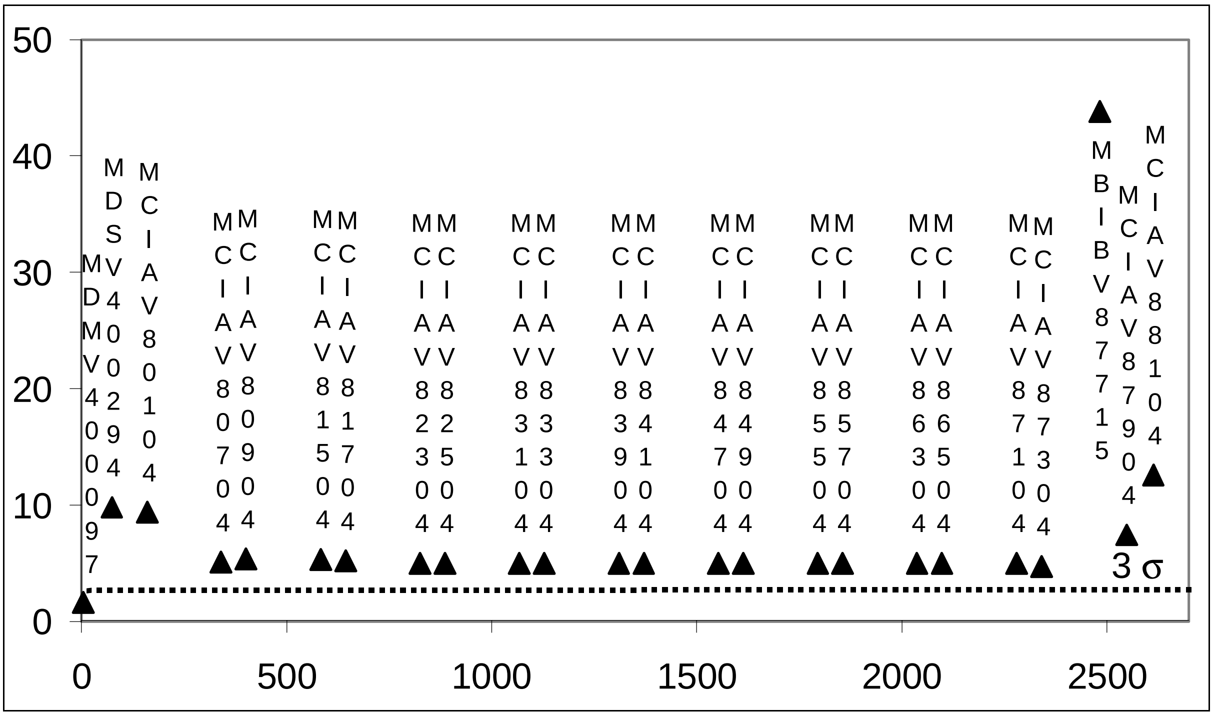


Figure 47: Corrector limits in error σ.

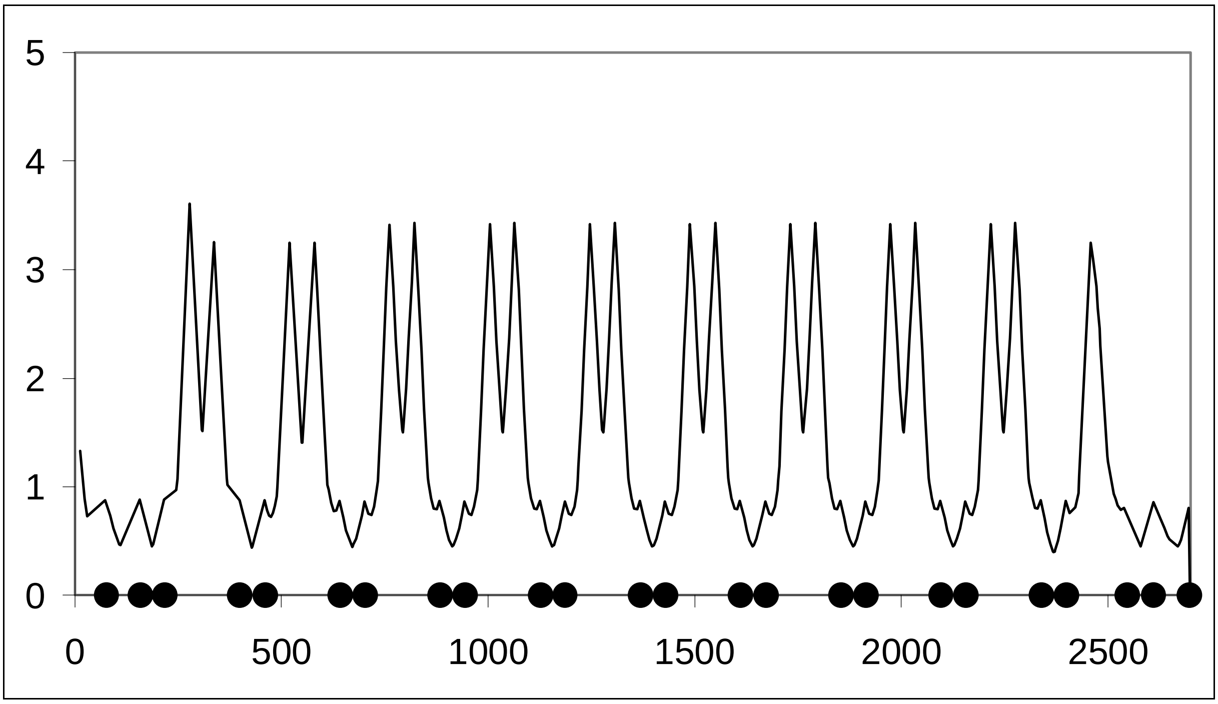


Figure 48: Resulting orbit error envelope in [mm] along TI 8 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 20 [μrad].

3.4.3 TI 8: 2-in-3 for monitors, 1-in-3 for correctors, horizontal plane

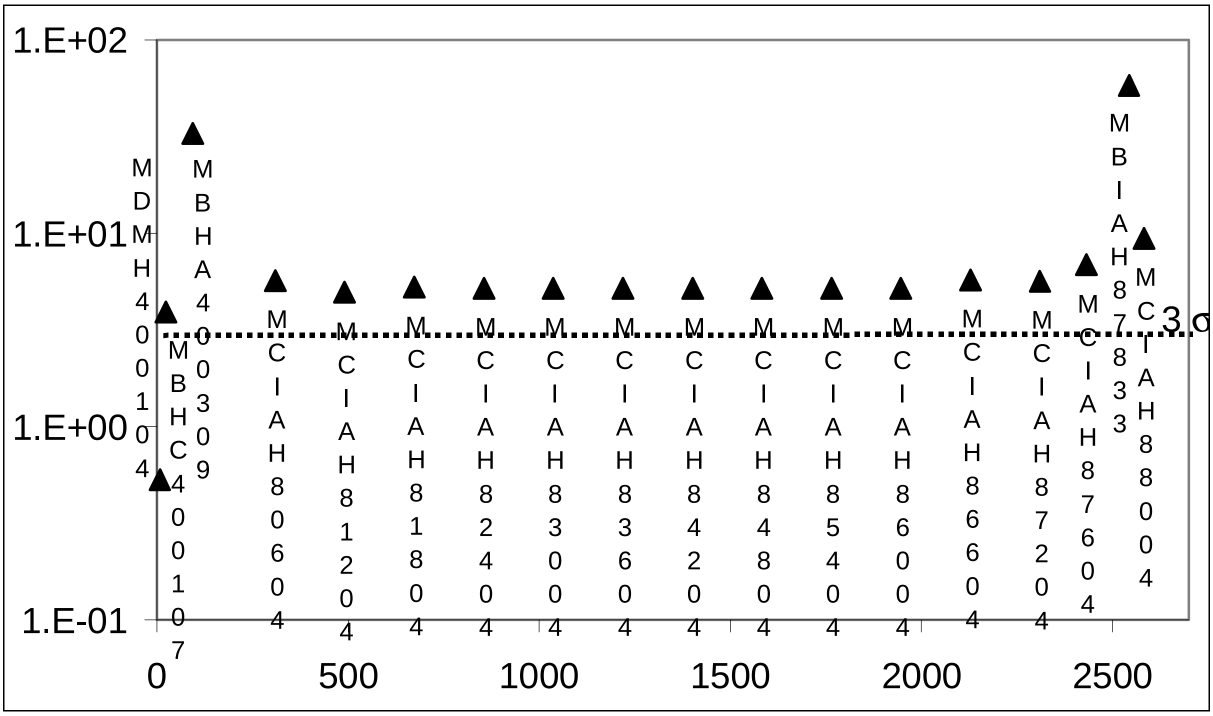


Figure 49: Corrector limits in error σ.

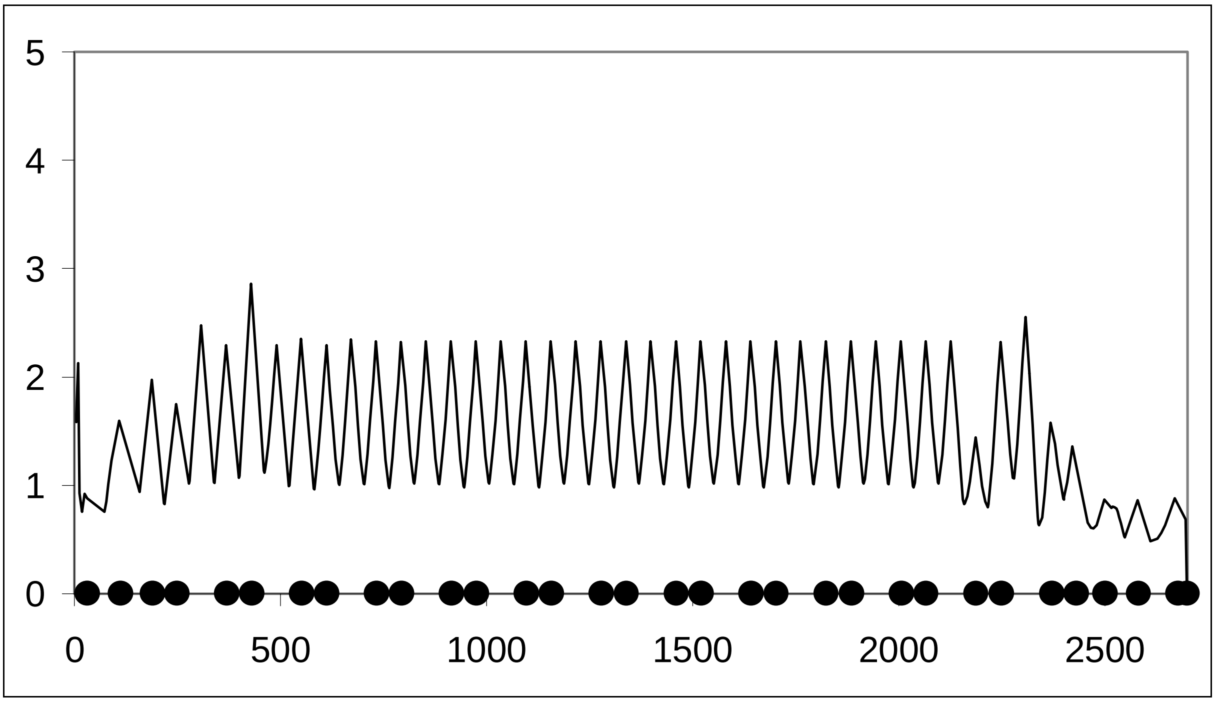


Figure 50: Resulting orbit error envelope in [mm] along TI 8 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 22 [μrad].

3.4.4 TI 8: 2-in-3 for monitors, 1-in-3 for correctors, vertical plane

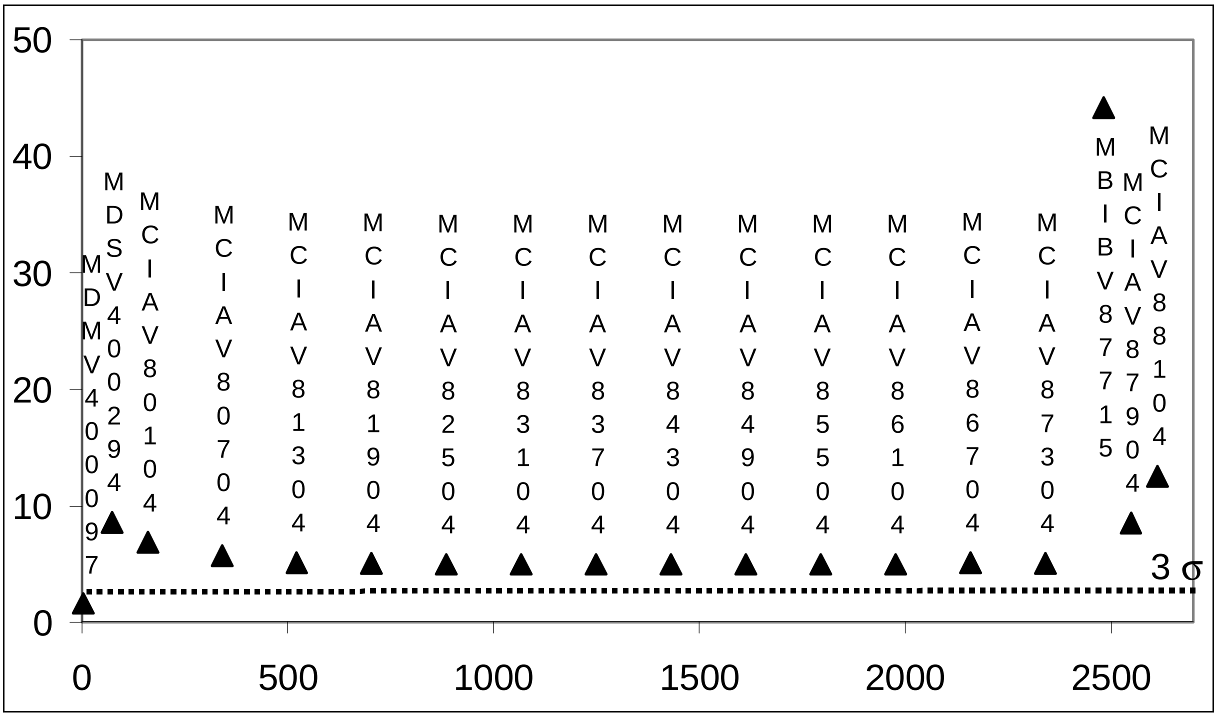


Figure 51: Corrector limits in error σ.

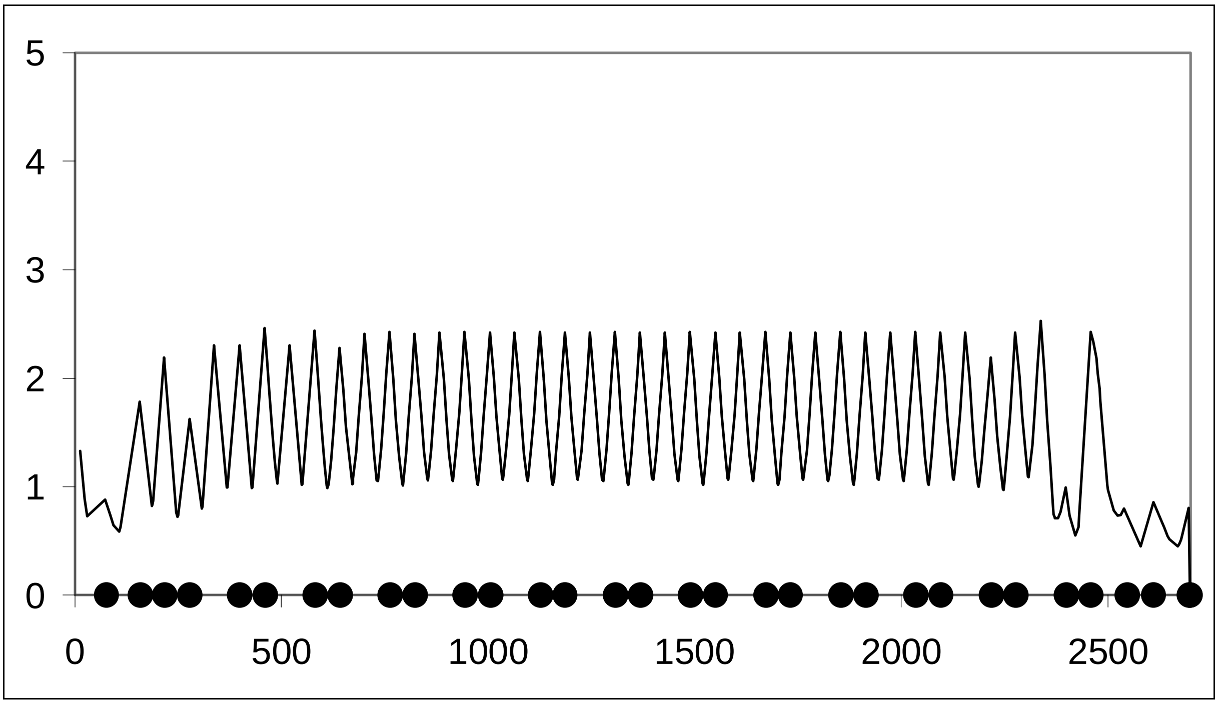


Figure 52: Resulting orbit error envelope in [mm] along TI 8 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 20 [μrad].

3.4.5 TI 8: 1-in-3 for monitors, 1-in-3 for correctors, horizontal plane

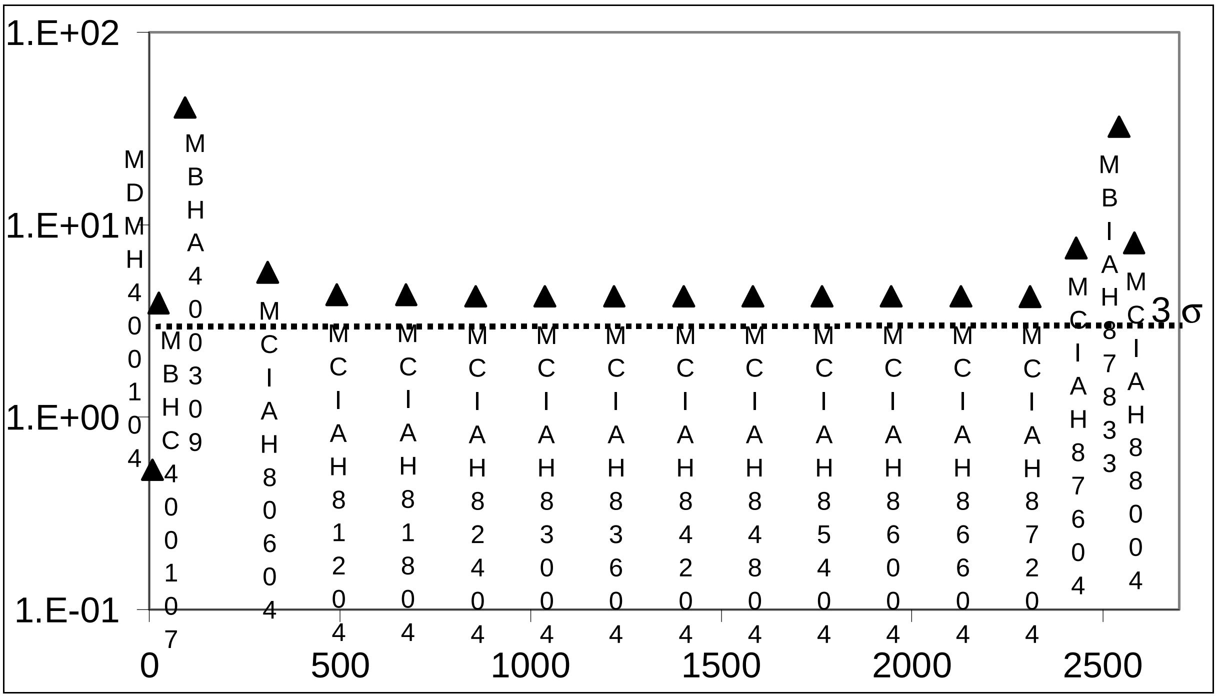


Figure 53: Corrector limits in error σ.

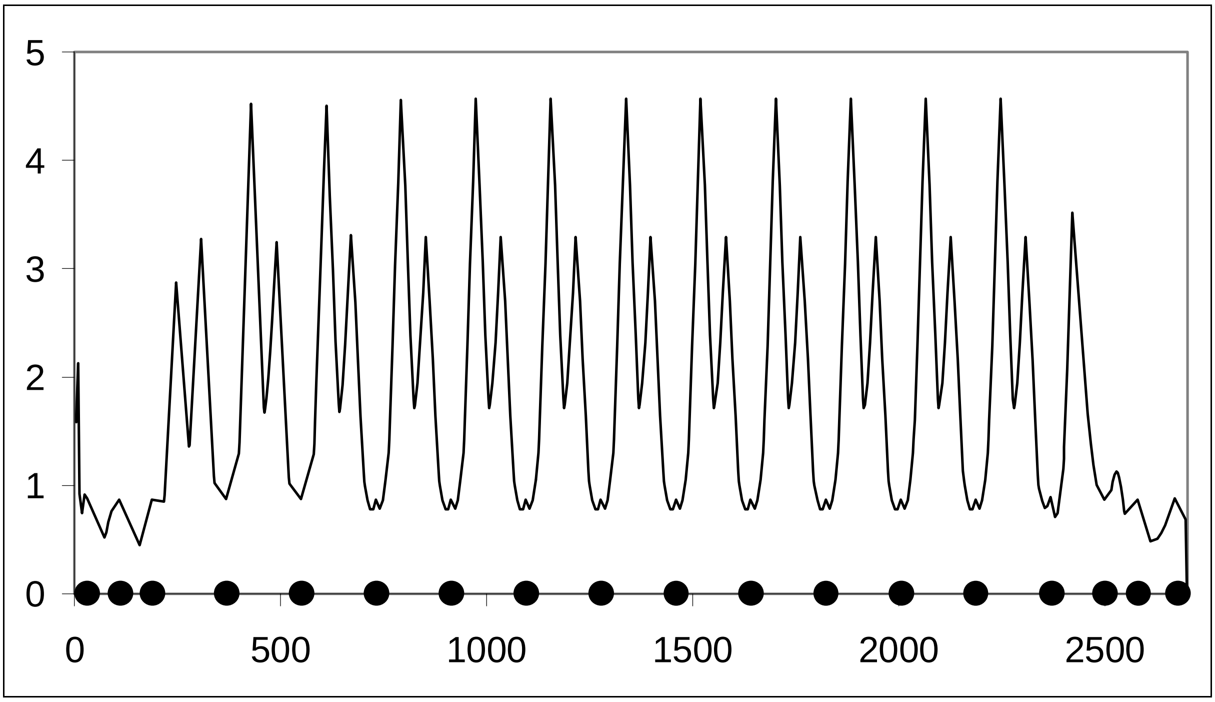


Figure 54: Resulting orbit error envelope in [mm] along TI 8 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 20 [µrad].

3.4.6 TI 8: 1-in-3 for monitors, 1-in-3 for correctors, vertical plane

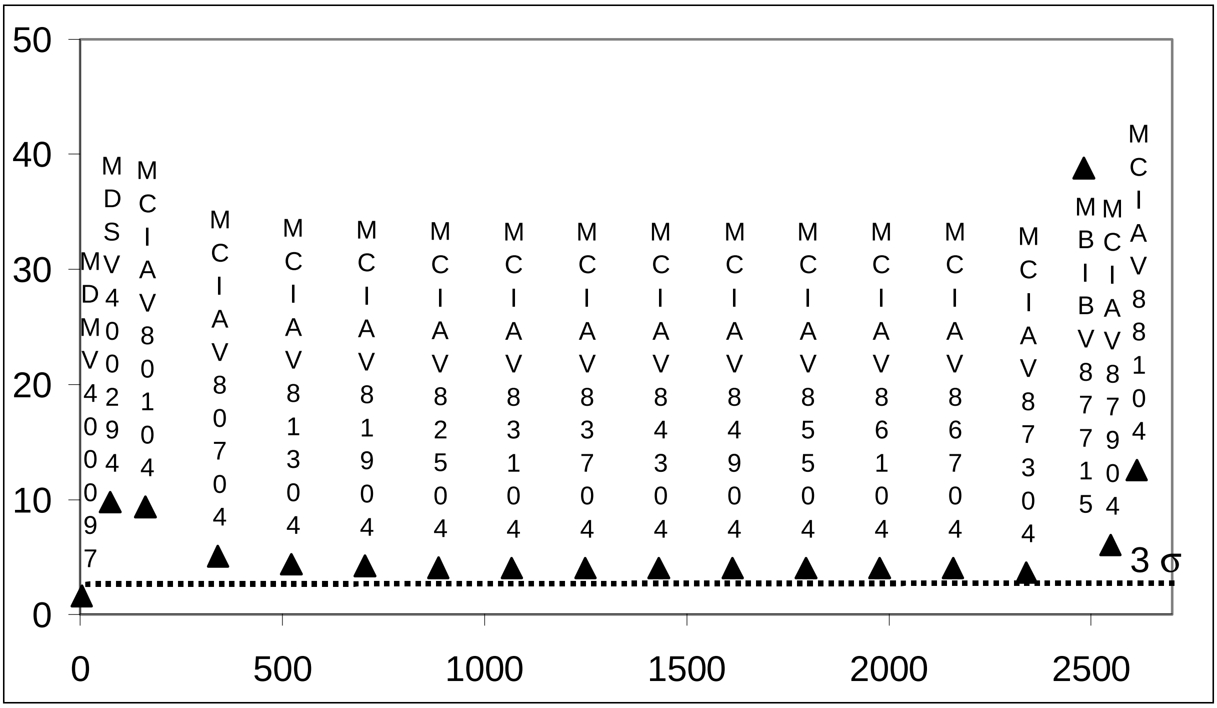


Figure 55: Corrector limits in error σ.

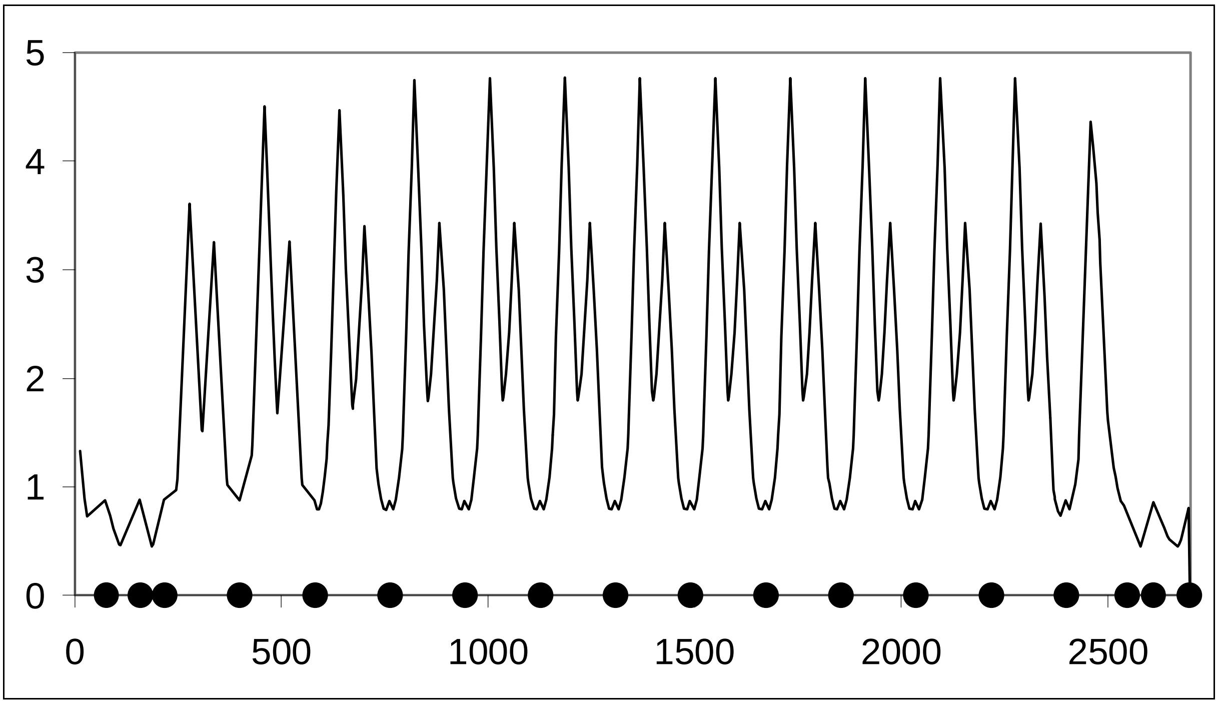


Figure 56: Resulting orbit error envelope in [mm] along TI 8 (in [m]; dots: monitor positions); 3 σ envelope value of the outgoing angle = 20 [μrad].

3.5 Overview of the Proposed Configurations

The overall element count and the maximum orbit envelope for the various optimised configurations are summarised in Table 4. Envelope values given in parentheses and denoted with '^' represent singular, very localised peak values.

The optimised 2-in-4 scheme requires, for TI 2 and TI 8 together, 90 new correctors (with a total of 102 corrective elements used). The same scheme in the previous study used a total of 110 corrective elements, of which 94 were assumed at that time to consist of modified LEP correctors. Maximum orbit excursions in the un-optimised scheme were also found to be around 3.5 mm.

Scheme	Corrective elements			Monitors			Maximum envelope [mm]	
	existing*	new	total	existing*	new	total	horizontal	vertical
TI 2 (incl. TT60)								
2-in-4	7	47	54	6	48	54	3.3	3.5
1-in-3	7	33	40	5	35	40	4.6	4.6
"2-in-3"	7	33	40	6	66	72	2.3 (^ 3.2, 3.2, 2.8)	2.4 (^ 3.0)
TI 8 (incl. TT40)								
2-in-4	5	43	48	2	46	48	3.3	3.4
1-in-3	5	31	36	2	34	36	4.5	4.7
"2-in-3"	5	31	36	2	60	62	2.3 (^ 2.8, 2.5)	2.4

Table 4: Element count and maximum orbit envelope for the various optimised schemes.

Table 5 to Table 8 give the detailed corrector-monitor configurations of the proposed schemes. Elements with grey background denote existing elements, e.g., those installed in beam lines from which TI 2 and TI 8 branch off, those to be located in the LHC main ring, or those defining the line geometry (groups of dipoles). These are elements that can be conveniently used for trajectory correction. Elements of the last category are also marked in **boldface**. One element, marked in *italic*, concerns a location in TT60 where we recommend to complement the presently installed configuration with a new corrector for optimum results. Each line in the Tables stands for a corresponding corrector-monitor pair[12]. A second element, marked in *italic*, concerns one element in TT40 which has been proposed in addition to the previous reference scheme for this part of TI 8.

[12] This statement does not apply to the supplementary monitors in the 2-in-3 schemes (Table 6 and Table 8), to be used in computer-supported SVD-style correction. The placement of these monitors in their respective row is solely meant to reflect their sequence in the beam line.

TI 2 2-in-4 scheme

Index	Plane		Corrector	Monitor
	H	V		
1		*	MDAV610013	BPCL610312
2	*		MDLH610104	BPCL610211
3	*		MDLH610206	BPCL610340
4		*	MDLV610304	BPCK610539
5	*		*MDLH610337*	BPMIH20204
6		*	MCIAV20104	BPMIV20304
7	*		**MBB20150**	BPMIH20404
8		*	MCIAV20304	BPMIV20504
9	*		MCIAH20804	BPMIH21004
10		*	MCIAV20904	BPMIV21104
11	*		MCIAH21004	BPMIH21204
12		*	MCIAV21104	BPMIV21304
13	*		MCIAH21604	BPMIH21804
14		*	MCIAV21704	BPMIV21904
15	*		MCIAH21804	BPMIH22004
16		*	MCIAV21904	BPMIV22104
17	*		MCIAH22404	BPMIH22604
18		*	MCIAV22504	BPMIV22704
19	*		MCIAH22604	BPMIH22804
20		*	MCIAV22704	BPMIV22904
21	*		MCIAH23204	BPMIH23404
22		*	MCIAV23304	BPMIV23504
23	*		MCIAH23404	BPMIH23604
24		*	MCIAV23504	BPMIV23704
25	*		MCIAH24004	BPMIH24204
26		*	MCIAV24104	BPMIV24304
27	*		MCIAH24204	BPMIH24404
28		*	MCIAV24304	BPMIV24504
29	*		MCIAH24804	BPMIH25004
30		*	MCIAV24904	BPMIV25104
31	*		MCIAH25004	BPMIH25204
32		*	MCIAV25104	BPMIV25304
33	*		MCIAH25604	BPMIH25804
34		*	MCIAV25704	BPMIV25904
35	*		MCIAH25804	BPMIH26004
36		*	MCIAV25904	BPMIV26104
37	*		MCIAH26404	BPMIH26604
38		*	MCIAV26504	BPMIV26704
39	*		MCIAH26604	BPMIH26804
40		*	MCIAV26704	BPMIV26904
41	*		MCIAH27204	BPMIH27404
42		*	MCIAV27304	BPMIV27504
43	*		MCIAH27404	BPMIH27604
44		*	MCIAV27504	BPMIV27704
45	*		MCIAH28004	BPMIH28204
46		*	MCIAV28104	BPMIV28304
47	*		MCIAH28204	BPMIH28404
48		*	MCIAV28304	BPMIV28504
49	*		MCIAH28804	BPMIH29004
50		*	MCIAV28904	BPMIV29104
51	*		MCIAH29004	BPMIH29304
52		*	MCIAV29104	BPMIV29504
53	*		**MBIBH29314**	BPMYB.Q5.L2
54		*	MCIAV29504	BPMYB.Q4.L2

Table 5: Listing of correctors and monitors in TI 2 for the 2-in-4 scheme.

TI 2 1(2)-in-3 scheme

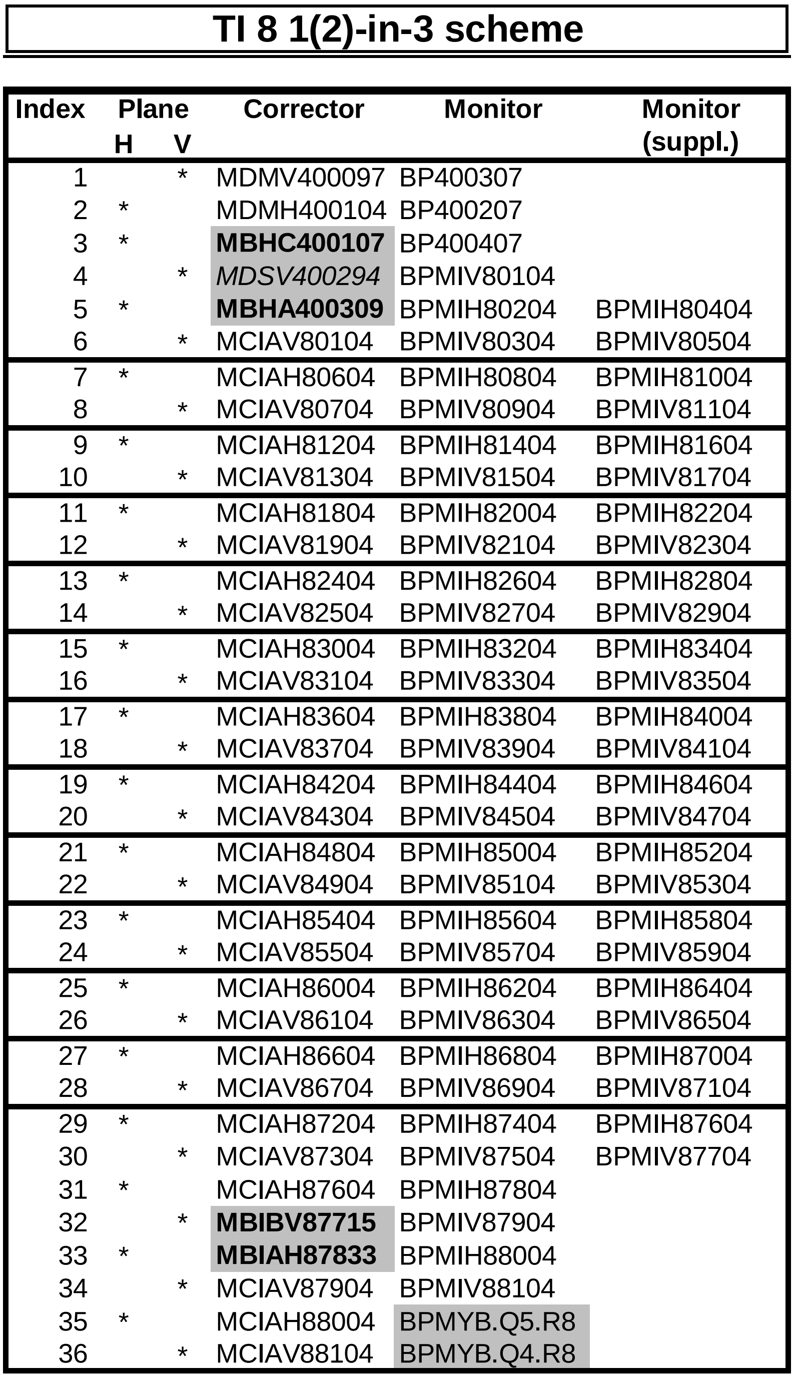
Index	Plane		Corrector	Monitor	Monitor
	H	V			(suppl.)
1		*	MDAV610013	BPCL610312	
2	*		MDLH610104	BPCL610211	
3	*		MDLH610206	BPCL610340	
4		*	MDLV610304	BPCK610539	
5	*		*MDLH610337*	BPMIH20204	
6		*	MCIAV20104	BPMIV20304	BPMIV20504
7	*		**MBB20150**	BPMIH20604	BPMIH20404
8		*	MCIAV20704	BPMIV20904	BPMIV21104
9	*		MCIAH20804	BPMIH21004	BPMIH21204
10		*	MCIAV21304	BPMIV21504	BPMIV21704
11	*		MCIAH21404	BPMIH21604	BPMIH21804
12		*	MCIAV21904	BPMIV22104	BPMIV22304
13	*		MCIAH22004	BPMIH22204	BPMIH22404
14		*	MCIAV22504	BPMIV22704	BPMIV22904
15	*		MCIAH22604	BPMIH22804	BPMIH23004
16		*	MCIAV23104	BPMIV23304	BPMIV23504
17	*		MCIAH23204	BPMIH23404	BPMIH23604
18		*	MCIAV23704	BPMIV23904	BPMIV24104
19	*		MCIAH23804	BPMIH24004	BPMIH24204
20		*	MCIAV24304	BPMIV24504	BPMIV24704
21	*		MCIAH24404	BPMIH24604	BPMIH24804
22		*	MCIAV24904	BPMIV25104	BPMIV25304
23	*		MCIAH25004	BPMIH25204	BPMIH25404
24		*	MCIAV25504	BPMIV25704	BPMIV25904
25	*		MCIAH25604	BPMIH25804	BPMIH26004
26		*	MCIAV26104	BPMIV26304	BPMIV26504
27	*		MCIAH26204	BPMIH26404	BPMIH26604
28		*	MCIAV26704	BPMIV26904	BPMIV27104
29	*		MCIAH26804	BPMIH27004	BPMIH27204
30		*	MCIAV27304	BPMIV27504	BPMIV27704
31	*		MCIAH27404	BPMIH27604	BPMIH27804
32		*	MCIAV27904	BPMIV28104	BPMIV28304
33	*		MCIAH28004	BPMIH28204	BPMIH28404
34		*	MCIAV28504	BPMIV28704	BPMIV28904
35	*		MCIAH28604	BPMIH28804	BPMIH29004
36		*	MCIAV28904	BPMIV29104	
37	*		MCIAH29004	BPMIH29304	
38		*	MCIAV29104	BPMIV29504	BPMIV29204
39	*		**MBIBH29314**	BPMIH29404	BPMYB.Q5.L2
40		*	MCIAV29504	BPMYB.Q4.L2	

Table 6: Listing of correctors and monitors in TI 2 for the 1(2)-in-3 schemes.

TI 8 2-in-4 scheme

Index	Plane		Corrector	Monitor
	H	V		
1		*	MDMV400097	BP400307
2	*		MDMH400104	BP400207
3	*		**MBHC400107**	BP400407
4		*	*MDSV400294*	BPMIV80104
5	*		**MBHA400309**	BPMIH80204
6		*	MCIAV80104	BPMIV80304
7	*		MCIAH80204	BPMIH80404
8		*	MCIAV80704	BPMIV80904
9	*		MCIAH80804	BPMIH81004
10		*	MCIAV80904	BPMIV81104
11	*		MCIAH81004	BPMIH81204
12		*	MCIAV81504	BPMIV81704
13	*		MCIAH81604	BPMIH81804
14		*	MCIAV81704	BPMIV81904
15	*		MCIAH81804	BPMIH82004
16		*	MCIAV82304	BPMIV82504
17	*		MCIAH82404	BPMIH82604
18		*	MCIAV82504	BPMIV82704
19	*		MCIAH82604	BPMIH82804
20		*	MCIAV83104	BPMIV83304
21	*		MCIAH83204	BPMIH83404
22		*	MCIAV83304	BPMIV83504
23	*		MCIAH83404	BPMIH83604
24		*	MCIAV83904	BPMIV84104
25	*		MCIAH84004	BPMIH84204
26		*	MCIAV84104	BPMIV84304
27	*		MCIAH84204	BPMIH84404
28		*	MCIAV84704	BPMIV84904
29	*		MCIAH84804	BPMIH85004
30		*	MCIAV84904	BPMIV85104
31	*		MCIAH85004	BPMIH85204
32		*	MCIAV85504	BPMIV85704
33	*		MCIAH85604	BPMIH85804
34		*	MCIAV85704	BPMIV85904
35	*		MCIAH85804	BPMIH86004
36		*	MCIAV86304	BPMIV86504
37	*		MCIAH86404	BPMIH86604
38		*	MCIAV86504	BPMIV86704
39	*		MCIAH86604	BPMIH86804
40		*	MCIAV87104	BPMIV87304
41	*		MCIAH87204	BPMIH87404
42		*	MCIAV87304	BPMIV87504
43	*		MCIAH87404	BPMIH87604
44		*	**MBIBV87715**	BPMIV87904
45	*		**MBIAH87833**	BPMIH88004
46		*	MCIAV87904	BPMIV88104
47	*		MCIAH88004	BPMYB.Q5.R8
48		*	MCIAV88104	BPMYB.Q4.R8

Table 7: Listing of correctors and monitors in TI 8 for the 2-in-4 scheme.



TI 8 1(2)-in-3 scheme

Index	Plane		Corrector	Monitor	Monitor
	H	V			(suppl.)
1		*	MDMV400097	BP400307	
2	*		MDMH400104	BP400207	
3	*		**MBHC400107**	BP400407	
4		*	*MDSV400294*	BPMIV80104	
5	*		**MBHA400309**	BPMIH80204	BPMIH80404
6		*	MCIAV80104	BPMIV80304	BPMIV80504
7	*		MCIAH80604	BPMIH80804	BPMIH81004
8		*	MCIAV80704	BPMIV80904	BPMIV81104
9	*		MCIAH81204	BPMIH81404	BPMIH81604
10		*	MCIAV81304	BPMIV81504	BPMIV81704
11	*		MCIAH81804	BPMIH82004	BPMIH82204
12		*	MCIAV81904	BPMIV82104	BPMIV82304
13	*		MCIAH82404	BPMIH82604	BPMIH82804
14		*	MCIAV82504	BPMIV82704	BPMIV82904
15	*		MCIAH83004	BPMIH83204	BPMIH83404
16		*	MCIAV83104	BPMIV83304	BPMIV83504
17	*		MCIAH83604	BPMIH83804	BPMIH84004
18		*	MCIAV83704	BPMIV83904	BPMIV84104
19	*		MCIAH84204	BPMIH84404	BPMIH84604
20		*	MCIAV84304	BPMIV84504	BPMIV84704
21	*		MCIAH84804	BPMIH85004	BPMIH85204
22		*	MCIAV84904	BPMIV85104	BPMIV85304
23	*		MCIAH85404	BPMIH85604	BPMIH85804
24		*	MCIAV85504	BPMIV85704	BPMIV85904
25	*		MCIAH86004	BPMIH86204	BPMIH86404
26		*	MCIAV86104	BPMIV86304	BPMIV86504
27	*		MCIAH86604	BPMIH86804	BPMIH87004
28		*	MCIAV86704	BPMIV86904	BPMIV87104
29	*		MCIAH87204	BPMIH87404	BPMIH87604
30		*	MCIAV87304	BPMIV87504	BPMIV87704
31	*		MCIAH87604	BPMIH87804	
32		*	**MBIBV87715**	BPMIV87904	
33	*		**MBIAH87833**	BPMIH88004	
34		*	MCIAV87904	BPMIV88104	
35	*		MCIAH88004	BPMYB.Q5.R8	
36		*	MCIAV88104	BPMYB.Q4.R8	

Table 8: Listing of correctors and monitors in TI 8 for the 1(2)-in-3 schemes.

3.6 Corrector Limits

From the results presented in the previous two sections it can be concluded that the predominant part of the correctors, including dipole groups, can correct orbit excursion caused by error distributions up to 3 σ. Exceptions are found in the correctors in both planes responsible for compensating injection errors, assumed to be Gaussian-distributed with RMS equal to 0.5 mm in position and 50 μrad in angle as described in section 2.4.2.

This deficit in corrector range becomes less severe, and eventually vanishes, as the injection error magnitudes are progressively reduced. A difference in the severity of such deficits can also be seen between the horizontal and vertical planes. In both TI 2 and TI 8, the vertical correctors responsible for injection-fix are advantageously located, both in terms of betatron amplitude and phase advance. A doubling of the range in these correctors will enable them to contain the assumed injection error with few adverse consequences. In the horizontal plane, on the other hand, optimal locations for correctors have to be found upstream of the injection point.

The horizontal correctors in all proposed schemes for injection-fix are located at compromised locations to mitigate, but not correct to within the specification, the injection error effects. The situation is further exacerbated by space constraints in this area. In any case it is not advisable to blindly increase the correction range of these horizontal correctors, because their less-than-optimal locations will result in excessive correction if they are forced to fix the injection errors. The rational solution should be strategically located horizontal corrector(s) in the extraction channel from the SPS to the TI 2 and TI 8 lines. Similar correctors in the vertical plane upstream would make it unnecessary to double the strengths of the vertical correctors for the injection-fix.

3.7 Analysis of Critical Elements

The method described in section 2.8 for identifying critical monitors and correctors, which uses as a measure the adverse impact on orbit error envelopes due to various failure modes, was applied to the configurations presented in sections 3.3 and 3.4. The five leading critical elements for each failure mode in each configuration are listed in Table 9 for TI 2 and Table 10 for TI 8. Apart from very few elements of relatively real concern, most of the large numbers seen in these Tables reflect artefacts of trivial configuration singularities or boundary effects, which can be summarised as follows:

In all 1-to-1 steering schemes, a missing monitor would immediately lead to a singular steering configuration and excessive correction can result if handled improperly. In such cases disabling one corrector would resolve the problem[13]. This applies also to the non-periodic sections of the over-constrained steering schemes, where the steering is actually 1-to-1. This can be demonstrated in the case of an over-constrained TI 8 line steering with BPMIH88204 in the horizontal or BPMIV80304 in the vertical plane missing. Table 10 shows orbit peaks of 586 mm and 686 mm respectively due to such singularities. By disabling the correctors MCIAH88404 and MCIAV80304, both peaks are reduced to around 3 mm. Similar solutions work for the 497-mm peak with missing BPMIV88304 in the vertical TI 8 steering, and the 106-mm peak with missing BPCL61340 in the horizontal TI 2 steering, as well as most of the other large numbers in the Tables due to missing monitors.

In the case of over-constrained steering in the periodic sections, the configurations are in general less susceptible to excessive corrections due to missing monitors. Minor susceptibility is seen in the vertical steering in TI 2, with a missing BPMIV28704 causing a 12-mm peak, and in TI 8, with a missing BPMIV87904 causing a 23-mm peak. The cause of

[13] A more effective algorithm for solving this problem can be found in [5].

these peaks is no longer the onset of singularity, but direct coupling between correctors and monitors almost 180° apart in phase. This problem can again be remedied by disabling correctors: MCIAV27704 in TI 2 and MCIAV88504 in TI 8, which reduced the orbit peaks to 3.5 mm and 4.5 mm respectively.

The large orbit peaks associated with missing BPMYB.Q5.L2(R8) in both TI 2 and TI 8 lines are only artefacts near the section boundary. As explained in section 2.4.3, these monitors, located beyond the end of TI 2 or TI 8, are used in the analysis to represent orbit-anchoring by monitors in the LHC injection area. The absence of such anchoring effects would undoubtedly create steering problems.

The large orbit peaks associated with individual monitor offset errors cannot be dismissed as artefacts, but should be regarded as potential sources of orbit correction problem. Prioritised preventive measures, such as BPM alignment and calibration, should be taken with these numbers in mind.

Almost all significant orbit peaks associated with missing correctors come from a single source: insufficient leverage in fixing injection errors. This problem is particularly severe in the horizontal plane, with a much smaller β_x than β_y. This has been obvious from the plots of corrector range in units of error σ's shown in the previous two sections. Both lines will be susceptible to this problem unless the solutions discussed in the previous section, namely horizontal orbit control upstream of the injection point, and enhanced vertical corrector range or upstream orbit control, are implemented. The corrector MCIAH80404, which is the leading offender in all other cases in TI 8 horizontal steering because of this injection error problem, does not appear in the list for missing correctors in the 1-in-3, 1-to-1 steering case. This is only because in this case the orbit peak caused by any one missing corrector in the periodic section is slightly (less than 1 %) larger than that caused by a missing MCIAH80404.

The limited correcting power at the beginning of the TI 2 and TI 8 lines for fighting injection errors is also responsible for the sensitivity of the orbit error to the correction process. This explains the large orbit peaks associated with corrector scaling errors being all caused by the same correctors responsible for fixing injection error. Again this sensitivity can be fixed only by orbit control upstream of the injection point.

The above discussion, especially that on monitor anomalies, helps illustrate the importance of an intelligent orbit correction algorithm. Independent of the degree of perfection of an orbit correction system[14], with monitors going out of service or presenting erratic data during operation, the difference between a competent algorithm and a questionable one can translate into a difference in orbit errors by a factor of 100 in the worst case. Actually the best way to deal with a missing monitor is not always to disable correctors, as suggested above, but to disable singular combinations of correctors [5], achievable only through an intelligent real-time algorithm.

3.8 Aperture Constraints

The analytical methods described in this report afford the possibility of further examining the performance of an orbit correction system in specific, localised areas. The entrance region into a horizontal dipole group towards the end of TI 8 required exactly such a close examination, due to the high β_y and higher-than-average η_y in this area, and the vertical aperture imposed by the horizontal dipole. The vertical underlying orbit error after correction was examined. It was found that, with all error sources taken into account, the 3 σ envelope was less than 1 mm in this localised region, within the stringent orbit excursion budget set

[14] Realistically speaking, making an orbit correction system too perfect can result in vulnerability to other problems, with monitor offset errors being the most intransigent example. Such problems can only be addressed through intelligent steering algorithms.

specifically for it. Under normal operating condition this “tight spot” should therefore not present a problem.

Disabled Monitor		**Fixed Monitor Offset of 3 mm**		**Disabled Corrector**		**Fixed Corrector Scale Error of 50 %**	
2-in-4 Monitor and Corrector (Horizontal)							
BPCL610211	9.65*	BPMIH20404	7.10	MDLH610104	10.18¶	MDLH610206	8.99¶
BPCL610340	7.17*	BPMIH28404	6.37	MDLH610206	6.96¶	MDLH610104	8.82¶
BPMIH29304	6.86*	BPMIH25204	6.22	MCIAH21004	4.47	MDLH610337	4.37
BPMYBQ5L2	5.20‡	BPMIH22004	6.21	MCIAH28204	4.01	MCIAH28204	4.21
BPMIH29004	4.84	BPMIH25804	6.21	MCIAH25004	3.96	MCIAH25004	4.15
2-in-4 Monitor and Corrector (Vertical)							
BPMIV29504	6.40*	BPCK610539	7.52	MDAV610013	17.65¶	MDAV610013	9.62¶
BPCL610312	4.63*	BPMIV20504	6.45	MCIAV24904	4.14	MCIAV24904	4.32
BPMIV25304	3.46	BPMIV25104	6.37	MCIAV24304	4.10	MCIAV24304	4.28
BPMIV24504	3.45	BPMIV23704	6.30	MCIAV24104	4.09	MCIAV24104	4.27
BPMIV25904	3.45	BPMIV24504	6.30	MCIAV23504	4.07	MCIAV23504	4.26
2-in-3 Monitor, 1-in-3 Corrector (Horizontal)							
BPCL610340	105.61*	BPCL610340	5.81	MDLH610104	10.18¶	MDLH610206	9.12¶
BPMYBQ5L2	10.53‡	BPMIH20204	4.49	MDLH610206	7.24¶	MDLH610104	8.94¶
BPCL610211	9.65*	BPMIH28204	4.15	MCIAH21404	4.69	MDLH610337	4.88
BPMIH28804	5.84	BPMIH20404	4.12	MCIAH23804	3.94	MCIAH21404	4.14
BPMIH22804	4.59	BPMIH21204	4.08	MCIAH24404	3.91	MCIAH22604	4.01
2-in-3 Monitor, 1-in-3 Corrector (Vertical)							
BPCK610539	19.06*	BPCK610539	7.74	MDAV610013	17.65¶	MDAV610013	9.81¶
BPMIV28704	11.74†	BPCL610312	4.82	MCIAV21304	4.51	MDLV610304	5.04
BPMIV29504	6.14†	BPMIV28104	4.02	MCIAV24304	4.08	MCIAV24304	4.17
BPCL610312	5.05*	BPMIV20304	3.89	MCIAV24904	4.07	MCIAV23704	4.16
BPMIV24504	4.76	BPMIV21104	3.88	MCIAV23704	4.07	MCIAV23104	4.13
1-in-3 Monitor and Corrector (Horizontal)							
BPCL610211	9.65*	BPMIH25804	7.51	MDLH610104	10.18¶	MDLH610206	9.86¶
BPCL610340	7.24*	BPMIH21604	7.49	MDLH610206	6.97¶	MDLH610104	8.96¶
BPMYBQ5L2	7.11‡	BPMIH24604	7.48	MCIAH21404	5.56	MCIAH25604	5.44
BPMIH28204	5.76*	BPMIH23404	7.48	MCIAH25604	5.30	MCIAH24404	5.43
BPMIH28804	5.60	BPMIH24004	7.48	MCIAH24404	5.30	MCIAH23804	5.43
1-in-3 Monitor and Corrector (Vertical)							
BPMIV29504	7.70*	BPMIV20304	7.82	MDAV610013	18.35¶	MDAV610013	9.62¶
BPMIV23904	4.76*	BPMIV25104	7.67	MCIAV21304	5.77	MCIAV24304	5.65
BPMIV24504	4.76*	BPMIV23304	7.67	MCIAV24304	5.52	MCIAV24904	5.64
BPMIV25104	4.76	BPMIV24504	7.67	MCIAV24904	5.51	MCIAV23704	5.63
BPMIV22704	4.75	BPMIV23904	7.67	MCIAV23704	5.50	MCIAV23104	5.58

*** Artefact caused by near singularity easily correctable by disabling correctors. See main text.**
† Artefact caused by phase anomaly, easily correctable by disabling correctors. See main text.
‡ Artefact caused by loss of anchoring point downstream. See main text.
¶ Artefact caused by insufficient leverage for correcting injection error. See main text.

Table 9: TI 2 peak 3 σ underlying orbit error envelope in [mm] due to element failure
(5 most critical elements listed for each failure mode).

3.9 Tilted Elements

A few dipoles in the beam lines are tilted, resulting in off-axes baseline co-ordinates in the local XY plane. This has no effect on the analysis presented in this report. The reason is that the analysis is entirely based on first order transfer matrices derived from the model, with no rotated quadrupoles, reflecting the actual alignment situation in the machine. Fixed co-ordinate change, even if off-axes, in the baseline trajectory caused by a tilted dipole is transparent to first order optics, and thus the analysis. Care must be given, on the other hand, to cases where monitors and correctors downstream of a tilted dipole are also rotated in accordance with the tilt, in which case mixing between the orbit error envelopes in the two planes at a constant combined probability density should be adopted instead. But this is not the case with the lines studied in this report.

Disabled Monitor		**Fixed Monitor Offset of 3 mm**		**Disabled Corrector**		**Fixed Corrector Scale Error of 50 %**	
2-in-4 Monitor and Corrector (Horizontal)							
BPMIH20404	19.84*	BPMIH27804	6.79	MDMH400104	4.73¶	MDMH400104	13.35¶
BPMIH20604	12.15*	BPMIH21404	6.18	MCIAH27604	4.30	MBHC400107	11.79¶
BPMIH28404	7.29*	BPMIH23204	6.17	MCIAH23004	3.94	MCIAH27604	4.41
BPMYBQ5R8	5.84‡	BPMIH24804	6.17	MCIAH25404	3.94	MCIAH26804	4.30
BPMIH20204	4.82	BPMIH25604	6.17	MCIAH24604	3.94	MCIAH26004	4.30
2-in-4 Monitor and Corrector (Vertical)							
BPMIV20304	6.04*	BPMIV20504	6.86	MCIAV20104	15.79¶	MCIAV20104	13.03¶
BPMIV27904	5.36*	BPMIV23104	6.31	MCIAV22904	4.11	MCIAV22904	4.29
BPMIV27704	3.65*	BPMIV25504	6.31	MCIAV25304	4.11	MCIAV23704	4.29
BPMIV23904	3.61	BPMIV23904	6.31	MCIAV24504	4.11	MCIAV25304	4.29
BPMIV22904	3.61	BPMIV22304	6.31	MCIAV26904	4.11	MCIAV26904	4.29
2-in-3 Monitor, 1-in-3 Corrector (Horizontal)							
BPMIH28204	585.68*	BPMIH27204	4.63	MDMH400104	4.80¶	MDMH400104	13.37¶
BPMIH20204	4.83	BPMIH20604	4.10	MCIAH21604	4.23	MBHC400107	12.10¶
BPMIH21404	4.68	BPMIH21404	4.03	MCIAH22804	3.95	MCIAH21604	4.12
BPMIH27204	4.64	BPMIH21804	3.82	MCIAH25204	3.92	MCIAH26404	4.01
BPMIH21804	4.62	BPMIH22604	3.82	MCIAH24604	3.92	MCIAH23404	4.01
2-in-3 Monitor, 1-in-3 Corrector (Vertical)							
BPMIV20304	685.92*	BPMIV27304	4.19	MCIAV20104	15.79¶	MCIAV20104	13.03¶
BPMIV28304	497.16*	BPMIV28104	3.94	MCIAV22904	4.12	MCIAV23504	4.18
BPMIV27904	22.76†	BPMIV22704	3.92	MCIAV21704	4.10	MCIAV25904	4.18
BPMIV27304	4.99	BPMIV21904	3.92	MCIAV25304	4.09	MCIAV24704	4.18
BPMIV21904	4.80	BPMIV26104	3.88	MCIAV24704	4.09	MCIAV24104	4.18
1-in-3 Monitor and Corrector (Horizontal)							
BPMIH20404	19.67*	BPMIH26604	7.48	MCIAH21604	5.38	MDMH400104	13.35¶
BPMIH20604	12.89*	BPMIH24204	7.48	MCIAH24004	5.30	MBHC400107	11.79
BPMIH28404	5.60*	BPMIH26004	7.48	MCIAH26404	5.30	MCIAH23404	5.44
BPMIH27204	5.50*	BPMIH23604	7.48	MCIAH25804	5.30	MCIAH26404	5.44
BPMIH20204	4.82	BPMIH24804	7.48	MCIAH23404	5.30	MCIAH25804	5.44
1-in-3 Monitor and Corrector (Vertical)							
BPMIV27904	6.84*	BPMIV24304	7.68	MCIAV20104	15.79¶	MCIAV20104	13.03¶
BPMIV20304	6.04*	BPMIV26704	7.68	MCIAV23504	5.53	MCIAV24104	5.66
BPMIV27304	4.83	BPMIV26104	7.68	MCIAV26504	5.53	MCIAV26504	5.66
BPMIV26704	4.77	BPMIV23704	7.68	MCIAV25904	5.53	MCIAV25904	5.66
BPMIV23704	4.77	BPMIV23104	7.68	MCIAV24104	5.53	MCIAV24704	5.66

*** Artefact caused by near singularity easily correctable by disabling correctors. See main text.**

† Artefact caused by phase anomaly, easily correctable by disabling correctors. See main text.

‡ Artefact caused by loss of anchoring point downstream. See main text.

¶ Artefact caused by insufficient leverage for correcting injection error. See main text.

Table 10: TI 8 peak 3 σ underlying orbit error envelope in [mm] due to element failure (5 most critical elements listed for each failure mode).

4 Conclusions

4.1 The Proposed Configurations for TI 2 and TI 8

The orbit correction configuration for the LHC injection transfer lines TI 2 and TI 8 have been studied using newly developed analytic methods. As a result, two candidate hardware configurations are proposed for each line. The first of the two is a 2-in-4 scheme capable of containing the nominal corrected orbit within envelopes of 3.3 - 3.5[15] mm at the 3 σ cutoff of total error distribution. It operates under a one-to-one (critically constrained) scenario. The second is a 2-in-3 monitor, 1-in-3 corrector scheme. It has two operating scenarios. One can use all the monitors and correctors in a (software-assisted) over-constrained steering to achieve evenly distributed 3 σ error-induced corrected orbit envelopes of 2.3 - 2.4 mm, with a few isolated peaks at no more than 3.2 mm. This can be the scenario for routine orbit correction once the lines are commissioned. Alternatively, during the commissioning phase, where intuitive one-to-one steering may be necessary, one can disable every other monitor in the periodic section to effectively perform such an operation, achieving 3 σ error-induced corrected orbit envelopes of 4.5 - 4.7 mm. The overall costs of the two proposed hardware configurations are comparable.

Besides the 3 σ error-induced corrected orbit envelope, many other performance criteria have been examined for each hardware configuration and steering scenario. The most important ones are as follows.

- Corrector range: The ability of the correctors (taking design specifications into account) to handle orbits caused by an overall error distribution cutoff at 3 σ has been examined. The analysis indicates that no design correction limits are encountered at a 3 σ error for the predominant majority of the correctors in TI 2 and TI 8. The only exceptions concern those cases where the assumed 3 σ injection errors pose strong demands on the first correctors. In the vertical plane this is rather a straightforward issue of strengths of the first correctors, already reasonably located, not being able to handle injection error extrema at 3 σ. Increasing their strengths appears to be the correct solution. In the horizontal plane, on the other hand, it is not possible to find advantageous locations within TI 2 / TI 8 proper for fixing injection errors. Indeed it is not recommended to force a very strong corrector in a poorly chosen location in this case. The solution must be found at a high β_x point upstream in the SPS-to-TI 2 / TI 8 extraction channels.
- Critical elements: The criterion of the 3 σ error-induced corrected orbit envelope is used to measure the impact of various element failures. These include monitors taken off-line or having excessive offsets, and correctors taken off-line or having excessive scaling errors. This analysis helps identify areas where particular attention must be paid to ensure stable operation. The leading critical elements in various failure modes are listed in Tables 9 and 10, of which the following are especially worth mentioning.
  - Singularity or phase anomaly: These adverse effects caused by missing monitors can be mitigated in all cases by proper disabling of correctors, for which an intelligent steering algorithm will be useful.
  - Monitor offset: The most serious cases can lead to non-trivial steering problems. Such problems have to be prevented mainly through hardware check, and can be mitigated to a smaller extent by an intelligent steering algorithm.

[15] See Table 7 for more details differentiating between planes and configurations.

— Insufficient leverage for injection-fix: This can result in considerable sensitivity to correctors at the beginning of the lines, either by their absence or errors in scaling. Apart from making sure that these very important correctors are functioning properly, the solutions proposed earlier to eliminate such deficiencies are also indispensable.

Overall, the two hardware configurations proposed for both TI 2 and TI 8 present different advantages in terms of the corrected orbit envelope. Their performances in terms of the other criteria are quite comparable, presuming the exceptions discussed above are correctly addressed.

4.2 The Analytic Method

The method developed for the current study has been presented in detail in this report. At the core this method relies on various generalised response matrices to completely characterise the performance of an orbit correction system based on inputs of optics and design specifications. The main advantage of this approach lies in its ability to present unambiguous answers in an efficient manner, which is useful when it comes to comparing different configurations, or improving existing ones. Its functionality can be grouped into the following categories.

- Performance criteria: These are various measures developed to quantify the observability, correctability, range, and singularity of a configuration, as well as its global behavior at arbitrary locations quantified in the form of various orbit and error envelopes. This provides a detailed and quantitative picture of the performance of a given configuration.
- Algorithms for configuration optimization: This can be invoked in two modes:
  — Analytic methods applied to detect and correct structural configuration defects based on quantitative criteria,
  — Numerical configuration fine-tuning used to optimize configurations where no structural defects are present and competing numerical factors make analytic comparison impractical.
  The second mode nonetheless relies on the analytic algorithms for measuring the performance of the intermediate configurations, without which efficient fine-tuning would be impossible.
- Other functions: The following methods are also developed:
  — Method to decompose a performance defect into its contributing error components, important for understanding the defect and finding its remedy,
  — Algorithm to identify critical elements and evaluate their impacts,
  — Efficient simulation taking advantage of the generalised response matrices in case results are needed that cannot be analytically evaluated.

The study of the TI 2 / TI 8 lines has been the first full-scale application of the expanded program, with nearly all the features applied. The physical properties and requirements of these lines in return provide useful reality checks and feature guidelines for the program itself. This exercise is thus beneficial to both endeavors. It should be noted that in scope the program is developed to deal with a generic orbit correction system. Application to other systems, in terms of both evaluation and optimization, should be straightforward[16].

[16] The program takes optics outputs from MAD, BeamOptics, or Optim. A suite of BeamOptics functions is being developed to set up the initial input templates and connect to the graphical displays of this program. An extension of this program to handle finite numbers of re-circulation and multiple lines with common elements has also been developed.

5 Acknowledgements

Eberhard Weisse is thanked for many helpful discussions. Malika Meddahi is thanked for helpful discussions and for preparing the configuration files for TI 8.

After the completion of this report the authors were informed by André Verdier of an earlier SL Division report by Volker Ziemann devoted to evaluating the effectiveness of the LHC orbit correction configuration. Although the scope and techniques of these two reports are quite different, readers interested in general techniques and different perspectives applicable to orbit correction configuration evaluation are encouraged to also consult [9]. André Verdier is also thanked for his other comments.

6 References

[1] A. Hilaire, V. Mertens, E. Weisse, “Beam Transfer to and Injection into LHC”, Proc. EPAC’98, Stockholm (1998), and LHC Project Report 208.

[2] A. Hilaire, V. Mertens, E. Weisse, “Trajectory Correction of the LHC Injection Transfer Lines TI 2 and TI 8”, LHC Project Report 122 (1997).

[3] G. de Rijk, B. Langenbeck (GSI Darmstadt), private communication.

[4] A. Hilaire, private communication.

[5] Y. Chao, “Orbit Correction Methods – Basic Formulation, Current Application at Jefferson Lab, and Future Possibilities”, Proc. Workshop on Automated Beam Steering and Shaping (ABS), Edited by M. Lindroos, CERN, (1998).

[6] P. Gill, W. Murray, M. Wright, “Practical Optimization”, Academic Press, 1981.

[7] W. Press, B. Flannery, S. Teukolsky, W. Vetterling, “Numerical Recipes in C”, Cambridge University Press, 1988.

[8] Y. Tong, “The Multivariate Normal Distribution”, Springer-Verlag, 1990[17].

[9] V. Ziemann, "On LHC Orbit Correction", CERN SL/93-51 (AP), 1993.

[17] In chapter 7 it is stated that one can only obtain bounds for such distributions if the covariance matrix E satisfies certain properties in its off-diagonal elements. The bound itself is not in analytical form, but obtainable through a numerical look-up table. This should be state-of-the-art as of 1990.

Appendices

A Focus of Mathematical Recipes in these Appendices

These appendices are mainly devoted to realise all calculations needed for solving the current set of problems through linear methods and recipes implementable through efficient and robust numerical algorithms. As mentioned in section 2, most calculations encountered here cannot be realistically carried out before being phrased in linear terms amenable to mature numerical algorithms, such as supported by *Mathematica*. Fortunately this is largely possible. We can enumerate this limited set of realistic operations:

- Matrix addition, multiplication, permutation, transpose, sub-matrices and direct sum (⊕);
- Inversion of non-degenerate square matrices;
- Matrix pseudo-inverse;
- Null space vectors;
- Eigenvalues and eigenvectors;
- Singular value decomposition (SVD).

Thus a detailed description is required of the recipes followed, from intuitive pictures to the input end of this limited set of algorithms, which is the main purpose of the following sections. Experience shows that these recipes allow us to reliably carry out calculations involving up to 750 dimensions (in a beam line consisting of over 1000 elements), in a time frame of less than 5 minutes for one complete set of analyses[18]. The same reliability and speed would be unthinkable if done through a non-linear solver or optimiser of *Mathematica*.

These recipes are given in enough detail that their translation into any programming language or numerical package is straightforward. For the current work *Mathematica* is employed to realise these recipes. The production version has been subjected to analytical and numerical tests to considerable length that led to our confidence in the outcome.

B Various Projection Operators

For a critically or over-constrained response matrix $\mathbf{M}^{\mathbf{CM}}$ the projection operators are given, as discussed in section 2.2, by

$$\begin{aligned} &\mathbf{\Pi}^{\parallel}_{\mathbf{CM}} = \mathbf{M}^{\mathbf{CM}} \cdot \mathbf{M}^{\dagger}_{\mathbf{CM}} = \mathbf{M}^{\mathbf{CM}} \cdot \left(\mathbf{M}^{\mathsf{T}}_{\mathbf{CM}} \cdot \mathbf{M}^{\mathbf{CM}}\right)^{-1} \cdot \mathbf{M}^{\mathsf{T}}_{\mathbf{CM}} \\ &\mathbf{\Pi}^{\perp}_{\mathbf{CM}} = \mathbf{I} - \mathbf{\Pi}^{\parallel}_{\mathbf{CM}} . \end{aligned} \tag{B.1}$$

In the special case where the response matrix $\mathbf{M}^{\mathbf{CM}}$ is rank-deficient such that even if the number of correctors is **larger** than that of the monitors, the corrector range still does not span the entire monitor space, we have to do this differently as follows

$$\begin{aligned} &\mathbf{\Pi}^{\perp}_{\mathbf{CM}} = \mathbf{O}^{\mathsf{T}}_{\mathbf{CM}} \cdot \mathbf{O}^{\mathbf{CM}} \ , \ \mathbf{\Pi}^{\parallel}_{\mathbf{CM}} = \mathbf{I} - \mathbf{\Pi}^{\perp}_{\mathbf{CM}} \\ &\mathbf{O}^{\mathbf{CM}} = \mathbf{N}^{\mathbf{CM}}_{\mathbf{null}}, \ \mathbf{N}^{\mathbf{CM}} = \mathbf{M}^{\dagger}_{\mathbf{CM}} \end{aligned} \tag{B.2}$$

where $\mathbf{M}^{\mathbf{CM}}_{\mathbf{null}}$ is as defined in section 2.

[18] This is timed on a Pentium III 600 MHz PC running *Mathematica 4.0*.

For under-constrained response matrices, the purpose of the projection operators is to divide up the **actuator** space, rather than the **responder** space, into a subspace that has an impact on the responders and one that does not. One way to express these operators is

$$\begin{aligned}
&\boldsymbol{\Pi}^{\parallel}_{\mathbf{CM}} = \mathbf{N}^{\mathbf{CM}} \cdot \mathbf{N}^{\dagger}_{\mathbf{CM}} = \mathbf{N}^{\mathbf{CM}} \cdot \left(\mathbf{N}^{\mathrm{T}}_{\mathbf{CM}} \cdot \mathbf{N}^{\mathbf{CM}}\right)^{-1} \cdot \mathbf{N}^{\mathrm{T}}_{\mathbf{CM}} \\
&\boldsymbol{\Pi}^{\perp}_{\mathbf{CM}} = \mathbf{I} - \boldsymbol{\Pi}^{\parallel}_{\mathbf{CM}} \\
&\mathbf{N}^{\mathbf{CM}} = \mathbf{M}^{\dagger}_{\mathbf{CM}}
\end{aligned} \tag{B.3}$$

satisfying

$$\begin{aligned}
&\mathbf{M}^{\mathbf{CM}} \cdot \boldsymbol{\Pi}^{\parallel}_{\mathbf{CM}} = \mathbf{M}^{\mathbf{CM}} \\
&\mathbf{M}^{\mathbf{CM}} \cdot \boldsymbol{\Pi}^{\perp}_{\mathbf{CM}} = \mathbf{0}.
\end{aligned} \tag{B.4}$$

One can also obtain them along the line of null space vectors, as must be clear from the last line in (B.4).

C Orthonormal Transformation and Decoupling of General Matrices

To find the orthonormal transformation that maps the existing vector space onto one spanned by a new basis consisting of two disjoint sets of unit vectors $\mathbf{U}^{\mathbf{I}}$ and $\mathbf{U}^{\mathbf{O}}$, representing respectively the subspaces including and excluding the row vectors of, for example, $\mathbf{M}^{\mathbf{EM}}$, the easiest approach is via the null space vectors

$$\begin{aligned}
&\mathbf{U}^{\mathbf{O}} = \mathbf{M}^{\mathbf{EM}}_{\mathbf{null}} \\
&\mathbf{U}^{\mathbf{I}} = \mathbf{U}^{\mathbf{O}}_{\mathbf{null}} \\
&\mathbf{T}^{\mathbf{EM}} = \mathbf{U}^{\mathbf{I}} \oplus \mathbf{U}^{\mathbf{O}}.
\end{aligned} \tag{C.1}$$

The symbol $\mathbf{T}^{\mathbf{EM}}$ in (C.1) represents the orthonormal transformation, a direct sum of the two matrices $\mathbf{U}^{\mathbf{I}}$ and $\mathbf{U}^{\mathbf{O}}$, which transforms the existing vector space into one where the first N basis vectors span the same subspace as the row vectors of $\mathbf{M}^{\mathbf{EM}}$, while the rest of the basis vectors are orthogonal to $\mathbf{M}^{\mathbf{EM}}$.

For convenience of later notation, we define the decoupling, via the above orthonormalisation, of any S×C mapping **N** from a C-dimensional space into the direct sum of two sub-matrices $\mathbf{N}^{\mathbf{M}}_{\mathrm{in}}$ and $\mathbf{N}^{\mathbf{M}}_{\mathrm{out}}$, respectively inside and outside the row-vector space of another R×C (C≥R) matrix **M** which shares the C-dimensional domain of **N**,

$$\begin{aligned}
\mathbf{N} \xrightarrow{\mathbf{T}^{\mathbf{M}}} \mathbf{N}' &= \mathbf{N} \cdot \left(\mathbf{T}^{\mathbf{M}}\right)^{\mathrm{T}} \\
&= \begin{pmatrix} \mathrm{N}'^{11} & \cdots & \mathrm{N}'^{1,\mathrm{R}} & \mathrm{N}'^{1,\mathrm{R}+1} & \cdots & \mathrm{N}'^{1,\mathrm{C}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{N}'^{\mathrm{S},1} & \cdots & \mathrm{N}'^{\mathrm{S},\mathrm{R}} & \mathrm{N}'^{\mathrm{S},\mathrm{R}+1} & \cdots & \mathrm{N}'^{\mathrm{S},\mathrm{C}} \end{pmatrix} \\
&= \mathbf{N}^{\mathbf{M}}_{\mathrm{in}} \oplus \mathbf{N}^{\mathbf{M}}_{\mathrm{out}}.
\end{aligned} \tag{C.2}$$

It can be easily verified that when **N**=**M**, thus R=S, $\mathbf{N}_{\mathrm{in}}^{\mathbf{M}}$ is invertable by construction and $\mathbf{N}_{\mathrm{out}}^{\mathbf{M}}$ is identically 0. In more complicated cases where **M** may be rank deficient, special care must be taken to ensure correct decoupling of these matrices.

D Projection of an Ellipsoid onto Lower and Higher Dimensions

We describe in the following how to obtain projections of an ellipsoid defined in (2.8) by

$$\mathrm{S} = \mathbf{A}^{\mathbf{E}^{\mathrm{T}}} \cdot \mathbf{E}^{\mathbf{E}} \cdot \mathbf{A}^{\mathbf{E}} \tag{D.1}$$

where S is the “radius” of the ellipsoid, to lower and higher dimensions via a transformation **M**. We discuss three possible cases below.

- Critically constrained

If M is critically constrained, namely square with full rank, then the mapping is trivially given by

$$\mathbf{E}^{\mathbf{E}} \rightarrow \mathbf{M}^{\mathbf{-1}^{\mathrm{T}}} \cdot \mathbf{E}^{\mathbf{E}} \cdot \mathbf{M}^{\mathbf{-1}}. \tag{D.2}$$

- Under-Constrained (Higher to Lower Dimension)

If M is under-constrained, namely the row rank Nr is lower than the column rank Nc[19], the projection is more involved because the ellipsoid is “collapsed” onto a lower dimensional space. The projection onto the lower space is defined by the “footprint” of the ellipsoid[20], reducing the Nc X Nc matrix E^{E} into an Nr X Nr matrix. To do this we form the set of Np=Nc-Nr equations

$$\begin{aligned} \nabla_{\mathbf{N}}^{\mathbf{i}} \mathrm{S} &= \mathbf{N}^{\mathbf{i}} \bullet \nabla \mathrm{S} \\ &= \sum_{\mathbf{kmn}} M_{\mathbf{null}}^{\mathbf{ik}} \frac{\partial A_{\mathbf{m}}^{\mathbf{E}} A_{\mathbf{n}}^{\mathbf{E}} E_{\mathbf{mn}}^{\mathbf{E}}}{\partial A_{\mathbf{k}}^{\mathbf{E}}} \\ &= 0, \qquad \mathbf{i} = \mathrm{Nr}+1,\ \mathrm{Nr}+2, \ldots \mathrm{Nc}, \end{aligned} \tag{D.3}$$

$$\mathbf{N} = \mathbf{M}_{\mathbf{null}}.$$

The geometric picture of (D.3) is that these equations define the points on the ellipsoid where the gradient vectors normal to the constant-radius contour are also normal to the null space vectors defined by **M**. Therefore, when these points are mapped along the null space vectors onto the lower dimension, their projection defines the outer boundary of the mapped ellipsoid. In principle (D.1) and (D.3) together already define the target ellipsoid in the lower dimension “back mapped” onto the original one.

In practice however, it is better to go a few extra steps to cast the problem in a solid matrix form. We can do this by first rotating the ellipsoid $\mathbf{E}^{\mathbf{E}}$ into a new basis by the transformation (C.1) such that in the new basis the first Nr basis are aligned with the image space of **M** (i.e., $\mathbf{U}^{\mathbf{I}}$ of (C.1)) and the last Np basis are aligned with the null space of **M**[21]. In

[19] This can happen even if the matrix is taller than wide.

[20] In this respect the assumption of convexity of the ellipsoid is crucial to the validity of our analysis.

[21] This ortho-normalisation is not essential but makes the picture clear.

the new basis the ellipsoid equation becomes $\mathbf{E'^E}$, and it is clear that the Np equations in (D.3) simply become the last Np rows of $\mathbf{E'^E}$. These equations together with (D.1), with $\mathbf{E^E}$ and $\mathbf{A^E}$ replaced by $\mathbf{E'^E}$ and $\mathbf{A'^E}$, define the projected ellipsoid "back mapped" onto the original ellipsoid in the new basis, which is an Nc-Np-1 (=Nr-1) dimensional object described by the intersection between the original ellipsoid and an Nr (=Nc-Np) dimensional hyper-plane. This suggests an elimination of the last Np variables in the ellipsoid equation since they are made dependent on the first Nr variables through the Np equations.

The outcome of this is the equation for the back-mapped sub-contour that directly corresponds to the mapped ellipsoid in the lower dimension. The elimination can be carried out using the Np equations (dropping the primes in **E** for simplicity)

$$\begin{pmatrix} E^{Nr+1,1} & \cdots & E^{Nr+1,Nr} & E^{Nr+1,Nr+1} & \cdots & E^{Nr+1,Nc} \\ E^{Nr+2,1} & \cdots & E^{Nr+2,Nr} & E^{Nr+2,Nr+1} & \cdots & E^{Nr+2,Nc} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ E^{Nc,1} & \cdots & E^{Nc,Nr} & E^{Nc,Nr+1} & \cdots & E^{Nc,Nc} \end{pmatrix} \bullet \begin{pmatrix} X^1 \\ \vdots \\ X^{Nr} \\ X^{Nr+1} \\ \vdots \\ X^{Nc} \end{pmatrix} \tag{D.4}$$

$$\begin{aligned} &= \left(E_R^E \oplus E_P^E\right) \bullet \left(X_R \oplus X_P\right) \\ &= E_R^E \bullet X_R + E_P^E \bullet X_P \\ &= 0 \end{aligned}$$

from which we can readily see that the elimination of the Np variables Xp in terms of Xr in the ellipsoid equation can be realised through

$$X_P = -\left(E_P^E\right)^{-1} \cdot E_R^E \cdot X_R = \mathbf{K} \cdot X_R$$

$$\mathbf{T}^{\mathbf{K}} = \mathbf{I} \oplus \mathbf{K} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 1 \\ K^{11} & K^{12} & \cdots & K^{1,Nr} \\ \vdots & \vdots & \ddots & \vdots \\ K^{Np,1} & K^{Np,2} & \cdots & K^{Np,Nr} \end{pmatrix} \tag{D.5}$$

and

$$\mathbf{E''^E} = \mathbf{T}^{\mathrm{T}} \cdot \mathbf{E'^E} \cdot \mathbf{T} \tag{D.6}$$

where we recovered the primes on $\mathbf{E'^E}$. The ellipsoid $\mathbf{E''^E}$ of equation (D.5) is a function of only Nr co-ordinates, which are coincident with the Nr co-ordinates in the lower dimension by construction. It follows that $\mathbf{E''^E}$ is the projection of $\mathbf{E'^E}$ onto the lower dimension, with the superfluous co-ordinates in the null space "squashed out" by **T** of (D.4). We need to remember that all this has been done in the ortho-normalised basis, so a trivial rotation back to the original basis is necessary.

- Over-constrained (Lower to Higher Dimension)

The need for projecting an ellipsoid into a higher dimensional space arises when we need to find the extreme values of certain operators in the higher dimensional space on ellipsoidal surfaces defined by lower dimensional distributions. The image of a lower dimensional ellipsoid by a map **M** (with numbers of rows and columns being Nr and Nc) into a higher dimension cannot be simply expressed by a symmetric matrix similar to (D.1), since it does not have enough rank and needs additional constraint equations. An alternative is to map it into a subspace of the higher space spanned by column vectors of **M**, which can be obtained more cleanly by rotating the higher space to align with this subspace as demonstrated earlier in (C.1). The map **M** into the rotated space becomes

$$\mathbf{M}^{\mathrm{T}} \to \mathbf{M}^{\mathrm{T}} \cdot \mathbf{T}^{\left(\mathrm{M}^{\mathrm{T}}\right)} \tag{D.7}$$

where all superscripts T denote matrix transpose. Taking only the first Nc rows[22] of **M** gives an invertable square matrix $\mathbf{M}^{\mathbf{P}}$, which allows us to apply the method for the critically constrained case

$$\mathbf{E}^{\mathrm{E}} \to \mathbf{M}^{\mathbf{p\text{-}1}^{\mathrm{T}}} \cdot \mathbf{E}^{\mathrm{E}} \cdot \mathbf{M}^{\mathbf{p\text{-}1}}. \tag{D.8}$$

Equation (D.8) gives the projection of the ellipsoid onto the subspace in the higher space. One can then take this projection, as well as the mapping between this subspace and the complete original un-rotated higher space, to explore various extreme values on this mapped ellipsoid.

E Inverse Projection of Point(s) onto an Ellipsoid

When performing the mapping **M** of an ellipsoid **E**, as described in (D.1), onto another lower or higher dimensional space, once a point of interest **X** (e.g., the point with the longest length or the largest component along an axis) is identified on the mapped ellipsoid, we often need to find the point(s) **Z** on the original ellipsoid that mapped into this point. This is a problem of inverse mapping and again the algorithm differs between over-constrained and under-constrained maps.

- Critically constrained

In this case **Z** is trivially related to **X** by

$$\mathbf{Z} = \mathbf{M}^{-1} \cdot \mathbf{X}. \tag{E.1}$$

- Under-Constrained (Higher to Lower Dimension)

Again this is more involved, with the row rank Nr lower than the column rank Nc of **M**. In fact for arbitrary **X** in the mapped ellipsoid there should be more than one **Z** in the original ellipsoid E that are mapped into **X**, unless **X** itself lies on the boundary of the mapped ellipsoid, in which case there is only one corresponding **Z**. Fortunately for our analysis, and because of convexity of the ellipsoid as a premise, all points of interest do lie on the boundary of the mapped ellipsoid.

[22] In fact the remainder of **M** must be all 0.

We observe that the point **Z** in the higher space is uniquely determined by N_c independent equations. These can be identified as follows. First we operate on **X** with the pseudo-inverse of **M**

$$\mathbf{Y}=\mathbf{M}^{\dagger}\cdot\mathbf{X} \tag{E.2}$$

Y is not necessarily the desired solution **Z** on the original ellipsoid **E**, but it is connected to **Z** by a combination of null space vectors of **M**. This means

$$\begin{aligned} &\mathrm{M}\bullet(\mathrm{Z}-\mathrm{Y})=0, \\ &\sum_{k}\mathrm{M}^{ik}\cdot\left(\mathrm{Z}^{k}-\mathrm{Y}^{k}\right)=0, \quad i=1,2,\ldots.\mathrm{Nr}. \end{aligned} \tag{E.3}$$

Also, the gradient vector at **Z** must be normal to the null space vectors of **M** so that its projection onto the lower dimension falls on the ellipsoid boundary. Thus

$$\begin{aligned} &\mathbf{N}=\mathbf{M}_{\mathbf{null}}, \\ &\nabla_{\mathbf{N}}^{\mathbf{i}}\mathrm{S}\Big|_{\mathbf{Z}}=2\,\mathbf{N}^{\mathbf{i}}\cdot\mathrm{E}\cdot\mathrm{Z}=0, \qquad i=\mathrm{Nr}+1,\,\mathrm{Nr}+2,\ldots\mathrm{Nc}. \end{aligned} \tag{E.4}$$

This is basically (D.3), and we also used (2.15). It is clear that (E.2), (E.3) and (E.4) provide a total of Nc independent equations sufficient to determine **Z**.

-Over-constrained (Lower to Higher Dimension)

This case is again somewhat trivial since **Z** is uniquely determined by **X** through the pseudo-inverse of **M**

$$\mathbf{Z}=\mathbf{M}^{\dagger}\cdot\mathbf{X}. \tag{E.5}$$

Many different **X**’s can lead to the same **Z**, but there is no ambiguity about the latter.

F Tangent Point between an Ellipsoid and an Arbitrary Hyper-Plane

The goal here is to find the “radius” S of an ellipsoid surface specified by the matrix **E** in an N-dimensional space defined by (D.1) at which it “inscribes”, or is tangent to, each face of a hyper-parallelogram[23] **P**. **P** is enclosed by Np pairs of hyper-planes, each consisting of two N-1 dimensional hyper-planes symmetric about the origin. These hyper-planes are represented by 2 X Np equations

$$\begin{aligned} &\sum_{\mathbf{k}}\mathbf{M}_{\mathbf{i}}^{\mathbf{k}}\cdot\mathbf{X}^{\mathbf{k}}=\mathbf{P}^{\mathbf{i}}(\mathbf{X})=\pm\mathbf{V}_{\mathbf{i}}, \\ &\nabla\,\mathbf{P}^{\mathbf{i}}=\sum_{\mathbf{k}}\hat{\mathbf{x}}^{\mathbf{k}}\mathbf{M}_{\mathbf{i}}^{\mathbf{k}}, \quad i=1,2,\ldots.\mathrm{Np}, \quad k=1,2,\ldots.\mathrm{N}, \end{aligned} \tag{F.1}$$

with the Np numbers $\mathbf{V_i}$’s providing a measure of the distances at which the hyper-planes are from the origin. In (F.1) we also showed that the gradient vector of the i-th hyper-plane is simply the vector with its components given by the elements of $\mathbf{M_i}$, the i-th row of **M**.

[23] The method discussed here should in fact work for more general cases not requiring symmetry of the hyper-polygon **P**, the only requirement for this method to be valid is that **P** be convex. We limit our model to the parallelogram here though, since that is all we encounter for the current analysis.

The ellipsoid and the i-th hyper-plane are tangent at points where their gradient vectors are parallel. From (2.15) and (F.1) this implies

$$2\mathbf{E}\cdot\mathbf{X}=\lambda_{\mathbf{i}}\,\mathbf{M}_{\mathbf{i}} \tag{F.2}$$

where $\lambda_{\mathbf{i}}$ is an undetermined scalar. It can be determined by taking (F.2) as an equation for **X** parameterised by $\lambda_{\mathbf{i}}$, which when substituted into (F.1) fixes the value of $\lambda_{\mathbf{i}}$, and thus the tangent point **X**, as a function of $\mathbf{V}_{\mathbf{i}}$. Finally the corresponding radius of the ellipsoid surface containing this tangent point can be trivially calculated by substituting **X** back into the ellipsoid equation.

In the special case where the parallelogram is a rectangle, i.e., all rows of **M** have only one non-zero element, the algebra described above is even further simplified.

This technique is used mainly in exploiting limits on correctability due to finite ranges imposed on correctors. One can find all the "radii" of the ellipsoid at which successive corrector ranges will be violated, and order the correctors by their "vulnerability". Thanks to convexity again in both the ellipsoid and the parallelogram, one never needs to worry about ambiguity in the ordering. That is, when we increase the radius of the ellipsoid from zero gradually and when at some point the ellipsoid touches a particular face of the parallelogram, it may be protruding out of another face even more already. But one can be absolutely sure that we did not miss that other face, the tangency to which must have happened earlier at a smaller radius. Thus the ordering of progressively less vulnerable correctors can be done unambiguously.

G Extreme Values of Arbitrary Operators on a Constrained Surface

We describe here the method for obtaining extreme values of the length of an operator on a constrained surface in N dimensions. To avoid complication due to exceptions caused by non-symmetry and non-convexity, we limit the treatment to ellipsoidal surfaces defined by **E** as in (D.1), which is all we need for the current analysis. The method itself is valid however for any constrained surfaces when exceptions are properly dealt with.

The operator of interest in N-dimension is of the form

$$\begin{aligned}&\Pi\cdot\mathbf{X}\mathbf{=}\mathbf{Y}\\&\mathbf{L}_{\Pi}=\mathrm{Y}^{\mathrm{T}}\cdot\mathrm{Y}=\mathrm{X}^{\mathrm{T}}\cdot\Pi^{\mathrm{T}}\cdot\Pi\cdot\mathrm{X}\end{aligned} \tag{G.1}$$

where **X** and **Y** are N-vectors and $\mathbf{L}_{\Pi}$ is a short hand for the length-squared of the product **Π** and **X**. The quantity $\mathbf{L}_{\Pi}$ now as a function of **X** is itself a scalar function in the N dimensional space with constant-value contours whose extreme value is what we seek to solve for, under the constraint of another scalar function for the ellipsoid as in (D.1). This is essentially an optimisation problem where the method of Lagrange multiplier can be useful. We first calculate the gradient of the scalar function $\mathbf{L}_{\Pi}$.

$$\begin{aligned}\nabla\mathrm{L}_{\Pi}&=\sum_{\mathbf{ijkm}}\hat{\mathbf{x}}_{\mathbf{i}}\frac{\partial\Pi_{\mathbf{jk}}X_{\mathbf{k}}\Pi_{\mathbf{jm}}X_{\mathbf{m}}}{\partial X_{\mathbf{i}}}=2\sum_{\mathbf{ijkm}}\hat{\mathbf{x}}_{\mathbf{i}}\delta_{\mathbf{ik}}\Pi_{\mathbf{jk}}\Pi_{\mathbf{jm}}X_{\mathbf{m}}\\&=2\sum_{\mathbf{ijm}}\Pi_{\mathbf{ji}}\Pi_{\mathbf{jm}}X_{\mathbf{m}}\hat{\mathbf{x}}_{\mathbf{i}}=2\Pi^{\mathbf{T}}\cdot\Pi\cdot\mathbf{X}\end{aligned} \tag{G.2}$$

The Lagrange multiplier method states that the extremum of $\mathbf{L}_{\Pi}$, constrained by the ellipsoid **E**, occurs when the two gradient vectors (G.2) and (2.15) are parallel, namely

$$\begin{aligned} &\mathbf{\Pi}^{\mathrm{T}}\cdot\mathbf{\Pi}\cdot\mathbf{X}=\lambda\,\mathbf{E}\cdot\mathbf{X}, \\ &\left(\mathbf{\Pi}^{\mathrm{T}}\cdot\mathbf{\Pi}\cdot\mathbf{E}^{-1}-\lambda\right)\cdot\mathbf{E}\cdot\mathbf{X}=\mathbf{0}. \end{aligned} \tag{G.3}$$

where λ is an undetermined scalar, and we have cast (G.3) in the form of an eigenvalue problem for the composite operator $\mathbf{\Pi}^{\mathrm{T}}\cdot\mathbf{\Pi}\cdot\mathbf{E}^{-1}$. Solving for the real eigenvectors in (G.3) immediately leads to the solution for **X** up to a scale factor. Imposing the “radius” S of the ellipsoid then uniquely determines the point where the extremum occurs.

Some examples of the operator **Π** help illustrate the applicability of this technique:

Distance from origin:	$\mathbf{\Pi}=\mathbf{I}$	
Projection onto the null space of $\mathbf{M}^{\mathbf{EM}}$:	$\mathbf{\Pi}=\mathbf{\Pi}_{\mathbf{EM}}^{\perp}$	
Component along the i-th axis:	$\mathbf{\Pi}=\hat{\mathbf{x}}_{\mathbf{i}}\cdot\hat{\mathbf{x}}_{\mathbf{i}}^{\mathrm{T}}$	
Component along the i-th axis inside null space of $\mathbf{M}^{\mathbf{EM}}$:	$\mathbf{\Pi}=\mathbf{\Pi}_{\mathbf{EM}}^{\perp}\cdot\hat{\mathbf{x}}_{\mathbf{i}}\cdot\hat{\mathbf{x}}_{\mathbf{i}}^{\mathrm{T}}$	(G.4)
Component along a vector **V**:	$\mathbf{\Pi}=\mathbf{V}\cdot\mathbf{V}^{\mathrm{T}}$	
Component along a vector **V** in higher dimension:	$\mathbf{\Pi}=\mathbf{K}\cdot\mathbf{V}\cdot\mathbf{V}^{\mathrm{T}}\cdot\mathbf{K}^{\mathrm{T}}$	

In the last example **K** is the matrix whose rows consist of the basis vectors in the lower space expressed in terms of the coordinates of the higher space under the (over-constrained) map.

We note that the number of real solutions of the eigenvalue problem depends on the rank of the operator **Π**. For example, the operator projecting out the component along the i-th axis is of rank 1, and can afford only one extremum, while more complicated operators can have a few <u>local</u> extrema on the ellipsoid **E**. In the latter case the global extremum is determined among the few local extrema trivially.

H Hessian and the Curvature of an Ellipsoid

Apart from the gradient type quantities such as (2.15), (D.3) and (E.4) for an ellipsoid, we can examine its second order behaviour at a given point. This is known as the Hessian of a matrix and roughly speaking gives a measure of the combined magnitude of the curvature around the point of interest.

$$\begin{aligned} \mathbf{H}_{\mathbf{E}}^{\mathbf{ij}}=\nabla_{\mathbf{i}}\nabla_{\mathbf{j}}\mathbf{E}^{\mathbf{E}}&=\sum_{\mathbf{km}}\frac{\partial^2 X_{\mathbf{k}}^{\mathbf{E}}X_{\mathbf{m}}^{\mathbf{E}}E_{\mathbf{km}}^{\mathbf{E}}}{\partial X_{\mathbf{i}}^{\mathbf{E}}\partial X_{\mathbf{j}}^{\mathbf{E}}}=2\sum_{\mathbf{km}}\delta_{\mathbf{ik}}\delta_{\mathbf{jm}}E_{\mathbf{km}}^{\mathbf{E}} \\ &=2\cdot\mathbf{E}^{\mathbf{E}}{}_{\mathbf{ij}}. \end{aligned} \tag{H.1}$$

Thus the Hessian of the ellipsoid is simply the ellipsoid matrix itself, due to its symmetry. It is a constant independent of the point of evaluation. The physical meaning of the Hessian is that it represents a measure of all the N second order derivatives of the constant-value contour along the N “principal axes” at any point. The Gram determinant (2.26) of $\mathbf{H}_{\mathbf{E}}$ is a <u>constant</u> under orthonormal transformations[24], and gives the product of all the second order derivatives

[24] This would be obvious if the ellipsoid matrix is diagonalised in (H.1).

of **E** in the basis where all the basis vectors are aligned along the principal axes of **E**. This property allows us to estimate how fast the value of an operator recedes from the extremum when moving away from it on the constant-probability ellipsoidal contour. If this number is large, indicating a very sharp tangent point between the ellipsoid and the operator contour defined by (G.2), then the extremum may not be as persistent as that corresponding to a small rate of recession of the operator value from the extremum, which indicates a "smooth" tangency. This "recession rate" from the extremum can be calculated as follows. We rotate the ellipsoid into a co-ordinate system such that one axis $\hat{\mathbf{V}}_{\mathbf{1}}$ is coincident with the gradient vector of **E** at the point of tangency, and the remaining axes are orthogonalised so as to make all cross terms in the Hessian zero[25]. The product of the second order derivatives along all the axes is, from (2.26)

$$\begin{aligned}\mathbf{G}_{\mathbf{H}_\mathbf{E}} &= \mathrm{Det}\left(\mathbf{H}_\mathbf{E}{}^{\mathrm{T}}\cdot\mathbf{H}_\mathbf{E}\right) = 4\,\mathrm{Det}\left(\mathbf{E}^{\mathrm{T}}\cdot\mathbf{E}\right)\\ &= \left(\prod_{\mathbf{j=1}}^{\mathbf{N}}\mathrm{S}_{\mathbf{j}}^{\mathbf{H}_\mathbf{E}}\right)^2 = \mathrm{S}_{\mathbf{1}}^{\mathbf{H}_\mathbf{E}\,2}\left(\prod_{\mathbf{j=2}}^{\mathbf{N}}\mathrm{S}_{\mathbf{j}}^{\mathbf{H}_\mathbf{E}}\right)^2\\ &= \left(\sum_{\mathrm{km}}^{\mathrm{N}}\frac{\partial^2 X_{\mathrm{k}}^{\mathrm{E}} X_{\mathrm{m}}^{\mathrm{E}} E_{\mathrm{km}}^{\mathrm{E}}}{\partial V_{\mathbf{1}}^{\mathrm{E}2}}\right)^2\left(\prod_{\mathbf{j=2}}^{\mathbf{N}}\sum_{\mathrm{km}}^{\mathrm{N}}\frac{\partial^2 X_{\mathrm{k}}^{\mathrm{E}} X_{\mathrm{m}}^{\mathrm{E}} E_{\mathrm{km}}^{\mathrm{E}}}{\partial V_{\mathbf{j}}^{\mathrm{E}2}}\right)^2\\ &= \left(\nabla_{\mathrm{V}_1}^2\mathrm{S}\right)^2\tilde{\mathbf{G}}_{\mathbf{H}_\mathbf{E}}\,,\end{aligned} \tag{H.2}$$

where we have singled out the first basis vector $\hat{\mathbf{V}}_{\mathbf{1}}$, made distinction between the original coordinates $\mathrm{X_i}$ and the rotated coordinates $\mathrm{V_i}$, and factored the easily calculated quantity $\mathbf{G}_{\mathbf{H}_\mathbf{E}}$ into the product of $\tilde{\mathbf{G}}_{\mathbf{H}_\mathbf{E}}$ and another easily calculated quantity

$$\begin{aligned}\nabla_{\mathrm{V}_1}^2\mathrm{S} &= \sum_{\mathrm{ijkm}}^{\mathrm{N}}\frac{\partial^2 X_{\mathrm{k}}^{\mathrm{E}} X_{\mathrm{m}}^{\mathrm{E}} E_{\mathrm{km}}^{\mathrm{E}}}{\partial X_{\mathbf{i}}^{\mathrm{E}}\partial X_{\mathbf{j}}^{\mathrm{E}}} V_{\mathbf{i}}^1 V_{\mathbf{j}}^1\\ &= \frac{1}{\left|\mathrm{V}\right|^2}\sum_{\mathrm{ijkmpq}}^{\mathrm{N}}\frac{\partial^2 X_{\mathrm{k}}^{\mathrm{E}} X_{\mathrm{m}}^{\mathrm{E}} E_{\mathrm{km}}^{\mathrm{E}}}{\partial X_{\mathbf{i}}^{\mathrm{E}}\partial X_{\mathbf{j}}^{\mathrm{E}}} E_{\mathbf{ip}}^{\mathrm{E}} X_{\mathbf{p}}^{\mathrm{E}} E_{\mathbf{jq}}^{\mathrm{E}} X_{\mathbf{q}}^{\mathrm{E}}\\ &= \frac{2}{\left|\mathrm{V}\right|^2}\sum_{\mathrm{ijkmpq}}^{\mathrm{N}}\delta_{\mathrm{ik}}\delta_{\mathrm{jm}} E_{\mathrm{km}}^{\mathrm{E}} E_{\mathbf{ip}}^{\mathrm{E}} X_{\mathbf{p}}^{\mathrm{E}} E_{\mathbf{jq}}^{\mathrm{E}} X_{\mathbf{q}}^{\mathrm{E}}\\ &= \frac{2}{\left|\mathrm{V}\right|^2}\mathbf{X}^{\mathbf{T}}\cdot\mathbf{E}^{\mathbf{T}}\cdot\mathbf{E}\cdot\mathbf{E}\cdot\mathbf{X} = \frac{2}{\left|\mathrm{V}\right|^2}\mathbf{X}^{\mathbf{T}}\cdot\mathbf{E}^{3}\cdot\mathbf{X}.\end{aligned} \tag{H.3}$$

where

$$\left|\mathrm{V}\right|^2 = \mathbf{V}^{\mathbf{1}\mathbf{T}}\cdot\mathbf{V}^{\mathbf{1}} = \mathbf{X}^{\mathbf{T}}\cdot\mathbf{E}^{\mathbf{T}}\cdot\mathbf{E}\cdot\mathbf{X} = \mathbf{X}^{\mathbf{T}}\cdot\mathbf{E}^{2}\cdot\mathbf{X}\,, \tag{H.4}$$

and we used the fact that $\mathbf{V^1}$ is the gradient vector of **E,** and repeatedly the symmetry of **E**. It then follows from (H.2) and (H.3) that if $\tilde{\mathbf{G}}_{\mathbf{H}_\mathbf{E}}$ is the measure of the "recession rate" from the extremum, it can be calculated in a straightforward manner. A large $\tilde{\mathbf{G}}_{\mathbf{H}_\mathbf{E}}$ corresponds to a less persistent extremum.

[25] This can always be done locally for a point on the ellipsoid.

I Issues Specific to Multi-Dimensional Gaussian Distribution

- Probability Density Distribution in the Mapped Space

Here we set out to confirm the relations given in (2.21). We first impose a trivial normalisation procedure on the map M such that we are free to assume that all σ_i's in (2.20) are 1. This is done by

$$\begin{gathered} \mathrm{x} \rightarrow \mathrm{C} \cdot \mathrm{x}, \\ \mathrm{M} \rightarrow \mathrm{M} \cdot \mathrm{C}^{-1}, \\ \mathrm{y} = \mathrm{M} \cdot \mathrm{x} \rightarrow \mathrm{M} \cdot \mathrm{C}^{-1} \cdot \mathrm{C} \cdot \mathrm{x} = \mathrm{y}, \\ \mathrm{C} = \begin{pmatrix} 1/\sigma_1 & 0 & 0 & 0 \\ 0 & 1/\sigma_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1/\sigma_\mathrm{N} \end{pmatrix}. \end{gathered} \tag{I.1}$$

In other words, we express the original coordinates in units of their respective σ's, and re-scale **M** accordingly. Thus the ellipsoid matrix **E** in this new coordinate system is just the identity matrix, and the contour of constant probability density a sphere of radius S. From the third line of (I.1) we see that this change of variable does not affect the image space, nor its probability distribution.

Thus under the (new) map **M**, the (new) ellipsoid **E** is mapped into the same twisted ellipsoid **F** as before the transformation (I.1) took effect:

$$\mathrm{x}^\mathrm{T} \cdot \mathrm{E} \cdot \mathrm{x} \xrightarrow{\mathrm{M}} \mathrm{y}^\mathrm{T} \cdot \mathrm{F} \cdot \mathrm{y} \tag{I.2}$$

with **F** obtained through one of the algorithms for ellipsoid projection discussed earlier depending on whether **M** is under, critically, or over-constrained. It will most likely develop off-diagonal elements as opposed to **E**. If we look along an axis $\hat{\mathrm{y}}_\mathrm{j}$ in the image space **Y** and ask for the probability distribution along this axis, we need to integrate out all the dimensions orthogonal to it. Namely,

$$\begin{aligned} P(y_\mathrm{j}) &= \int_{-\infty}^{\infty} dy_1 \ldots \int_{-\infty}^{\infty} dy_{\mathrm{j-1}} \int_{-\infty}^{\infty} dy_{\mathrm{j+1}} \ldots \int_{-\infty}^{\infty} dy_\mathrm{N} \, P(y_1, y_2, \ldots y_\mathrm{N}) \\ &= \mathrm{G} \int_{-\infty}^{\infty} dy_1 \ldots \int_{-\infty}^{\infty} dy_{\mathrm{j-1}} \int_{-\infty}^{\infty} dy_{\mathrm{j+1}} \ldots \int_{-\infty}^{\infty} dy_\mathrm{N} \, e^{-\mathrm{y}^\mathrm{T} \cdot \mathrm{F} \cdot \mathrm{y}}, \end{aligned} \tag{I.3}$$

where G is a normalising factor, retained since the map **M** from **x** to **y** will change the measure of the normalisation integral.

It is in fact easier to calculate this quantity in the original space of **x**. We notice that for a particular value of $\mathbf{y_j} = a$, its probability density comes from corresponding set of points in the **x**-space that was mapped by **M** so as to have the j-th component of its image equal to *a*.

$$\mathbf{M^j} \bullet \mathbf{x} = \sum_\mathbf{k} \mathbf{M^{jk}} \cdot \mathbf{x_k} = a. \tag{I.4}$$

This set of points is determined by two constraints: the null space vectors in the **x**-space for the j-th row of **M**, and any point in the **x**-space whose image under **M** gives the point on the j-th axis in the **y**-space at a distance *a* from the origin. The second constraint can be taken, for example, to be the pseudo-inverse of the j-th row of **M** acting on the point $\hat{\mathbf{a}}$ = (0,…0, *a*, 0,…0) in **y**-space. Thus the equation for this set of points is

$$\begin{aligned}&\left(\mathbf{x}-\mathbf{M}^{\dagger}\cdot\hat{\mathbf{a}}\right)=\left(\mathbf{N}^{\mathbf{M}^{\mathbf{j}}}_{\mathbf{null}}\right)^{\mathrm{T}}\cdot\mathbf{P},\\&\mathbf{P}=\left(p_1,p_2,\ldots,p_{N-1}\right),\end{aligned}\tag{I.5}$$

which basically states that **x** is connected to the pseudo-inverse (back mapped image) of $\hat{\mathbf{a}}$ through a linear combination of the N-1 null space vectors of the j-th row of **M**. Thus this point set is simply an N-1 dimensional hyper-plane in the x-space. The probability *P(y$_j$=a)* of (I.3) can be alternatively evaluated using the corresponding point set in the x-space defined in (I.5). The advantage of the latter is that, since we have normalised the co-ordinates in the x-space to make the probability contour a sphere, integration of probability density on the N-1 dimensional hyper-plane depends only on the distance of the plane from the origin, and not its orientation. From (I.4) it is clear this distance for the point set (I.5) is

$$\mathbf{D}^{\mathbf{j}}=\frac{a}{\left|\mathbf{M}^{\mathbf{j}}\right|},\tag{I.6}$$

where the denominator is simply the length of the j-th row of **M**. Integrating the probability density of a normalised spherical Gaussian distribution over such a hyper plane can be trivially done in the **x**-space, if without the loss of generality we rotate the coordinates so that the hyper plane is perpendicular to the first axis **x$_1$**:

$$\begin{aligned}P\left(y_{\mathrm{j}}=a\right)&=\frac{\mathrm{G}}{\sqrt{\pi}^{\mathrm{N}}}e^{-a^2/\left|\mathbf{M}_{\mathbf{j}}\right|^2}\int_{-\infty}^{\infty}d\,x_2\;....\int_{-\infty}^{\infty}d\,x_{\mathrm{N}}\,e^{-\sum_{i=2}^{\mathrm{N}}x_i^2}\\&=\frac{\mathrm{G}}{\sqrt{\pi}}e^{-a^2/\left|\mathbf{M}_{\mathbf{j}}\right|^2}.\end{aligned}\tag{I.7}$$

G can be easily determined as

$$\begin{aligned}&\int_{-\infty}^{\infty}da\,P\left(a\right)=\frac{\mathrm{G}}{\sqrt{\pi}}\int_{-\infty}^{\infty}da\,e^{-a^2/\left|\mathbf{M}_{\mathbf{j}}\right|^2}=\mathrm{G}\cdot\left|\mathbf{M}_{\mathbf{j}}\right|=1,\\&\mathrm{G}=\frac{1}{\left|\mathbf{M}_{\mathbf{j}}\right|}.\end{aligned}\tag{I.8}$$

Thus from (I.7) and (I.8) the probability density along the j-th axis in the **y**-space is simply that given in (2.21). It is still a Gaussian but with its RMS value magnified by a factor of $\left|\mathbf{M}_{\mathbf{j}}\right|$, with a reduction in amplitude by the same factor.

Back in the original **x**-space on a spherical surface of constant probability density with radius *b*, or *b* times the σ (=1) of the single axis distribution, the maximum projection from any point **x** in this spherical surface onto the j-th axis in the **y**-space is easily seen to be

$$\tilde{\mathbf{x}} = b \cdot \frac{\mathbf{M^j}}{\left|\mathbf{M^j}\right|}, \quad \mathbf{y^j} = \mathbf{M^j} \cdot \tilde{\mathbf{x}},$$
$$\left|\mathbf{y^j}\right| = b \cdot \left|\mathbf{M^j}\right| = b \cdot \boldsymbol{\sigma}_\mathbf{y}^\mathbf{j}, \tag{I.9}$$

where $\boldsymbol{\sigma}_\mathbf{y}^\mathbf{j}$ is the RMS of the distribution along the j-th axis in the **y**-space, or simply $\left|\mathbf{M}_\mathbf{j}\right|$ by (I.7). Equation (I.9) states the important result that the maximal projection onto the **y**-space, from the distribution in the **x**-space contained within a contour corresponding to *b* times the single-axis RMS for each axis[26], is still *b* times the single-axis RMS for each axis after remaining dimensions are integrated out.

This result is especially powerful, thanks to the fact that the multi-dimensional Gaussian distribution can be factored, in that it is valid no matter whether **M** is critically, over, or under constrained. In some sense we only deal with the null space of one row of **M** at a time, which is always under-constrained. We should also emphasise that although (2.21) gives the per-axis distribution, the overall distribution as a function of all co-ordinates in the **y**-space is in general no longer the product of these distributions.

- Probability Density Distribution of Extrema and Length

In an N-dimensional Gaussian distribution one can look into two other types of distribution of physical significance and relevant to the current analysis. These are the distributions of overall extrema and length:

$$\begin{aligned} P_{max}(\mathbf{m}) &= P(max(x_1, x_2, \dots x_N)), \\ P_{min}(\mathbf{n}) &= P(min\ (x_1, x_2, \dots x_N)), \\ P_{length}(\mathbf{r}) &= P\left(\sqrt{\sum x_i^2}\right), \end{aligned} \tag{I.10}$$

subject to normalising conditions. Due to the symmetry present in all distributions encountered in the current analysis, it is more relevant to discuss the maximum of the absolute values than P_{max} and P_{min} of (I.10):

$$P_{|max|}(\mathbf{m}) = P(max(|x_1|, |x_2|, \dots |x_N|)). \tag{I.11}$$

For completely normalised and orthogonal N-dimensional Gaussian distributions, these can be evaluated in closed form as follows.

Distribution of Absolute Maximum:

We will focus on the definition of (I.11). This can be done, for a given value **m**, by forming a hyper-cube whose 2N faces are contained in the hyper-planes $x_1 = \pm\mathbf{m}$, $x_1 = \pm\mathbf{m}$, …. $x_N = \pm\mathbf{m}$. It is easy to see that if and only if a point lies on this hyper-cube would the maximum of the absolute values of all its co-ordinates be equal to **m**. This leads to the probability density distribution of **m**, which is simply the integration of the probability density over the cube surface for progressively larger cubes

[26] Of course ortho-normalisation is assumed in the **x**-space.

$$
\begin{aligned}
P_{|max|}(\mathbf{m}) &= \mathrm{G} \oint_{cube\,at\,\mathbf{m}} P(x_i) = \frac{\mathrm{G}}{\sqrt{\pi}^{\mathrm{N}}} \oint_{cube\,at\,\mathbf{m}} e^{-\sum_{i=1}^{\mathrm{N}} x_i^2} \\
&= \frac{2\mathrm{NG}}{\sqrt{\pi}^{\mathrm{N}}} e^{-\mathbf{m}^2} \prod_{i=1}^{\mathrm{N}-1} \int_{-\mathbf{m}}^{\mathbf{m}} e^{-x_i^2} d x_{\mathrm{i}} = \frac{2\mathrm{NG}}{\sqrt{\pi}} e^{-\mathbf{m}^2} \mathbf{erf}^{\mathrm{N}\text{-}1}(\mathbf{m}),
\end{aligned}
\tag{I.12}
$$

where we retained the normalising factor G and **erf** is the error function. This distribution, when N is large, leads to a shift of the peak away from the origin. However for orthogonal and normalised distributions this shift does not cause the majority of the distribution to lie much beyond 3 σ even for N=1000, as can be seen in Figure A.1. However, the equivalent of (I.12) for general correlated distributions, namely, distributions corresponding to an ellipsoid matrix with off-diagonal elements, cannot be computed in closed form [8]. An attempt to derive its possible range containing a given fraction of the total distribution, similar to the differentiation scheme to be discussed below for distributions of length, soon leads to intractable algebra. At this point for such general cases we can mostly rely on massive simulation, which shows that for most cases studied the 99% cutoff points for such distributions are at most 10 - 15 % beyond the 3 σ values. They also bear strong resemblance to the characteristics of Figure A.1 [2].

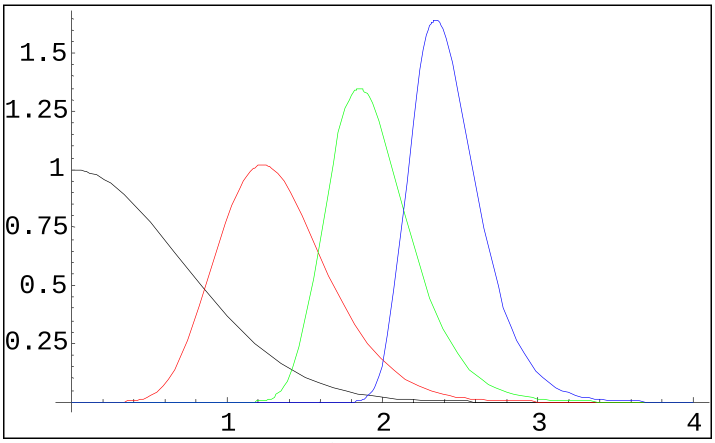


Figure A.1 Distribution of the absolute maximum in N dimensions (curves from left to right for N = 1, 10, 100 and 1000; abscissa: multiple of σ)

Distribution of Length:

Again, the distribution of the length of **x** can be done in closed form only for orthogonal distributions. This fact is however more relevant than in the case of distribution of the maximum, because we will use this closed form to perform certain cut-offs on the initial error distribution, which is orthogonal.

This distribution P_{length} in (I.10) then, in the orthonormal case, is computed by integrating over an (N-1)-sphere with radius **r**. The result is well known:

$$
\begin{aligned}
P_{length}(\mathbf{r}) &= \oint_{sphere\,at\,\mathbf{r}} P(x_i) = \frac{1}{\sqrt{\pi}^{\mathrm{N}}} \oint_{sphere\,at\,\mathbf{r}} e^{-\sum_{i=1}^{\mathrm{N}} x_i^2} \\
&= \frac{2}{\Gamma(\mathrm{N}/2)} \mathbf{r}^{\mathrm{N}\text{-}1} e^{-\mathbf{r}^2},
\end{aligned}
\tag{I.13}
$$

with the understanding that **r** now runs only in the positive axis. Figure A.2 shows such distributions for various values of N. It can be seen that unlike with P_{max}, at much smaller values of N the shift away from centre is already severe for P_{length}.

Cutoff on Distribution of Length:

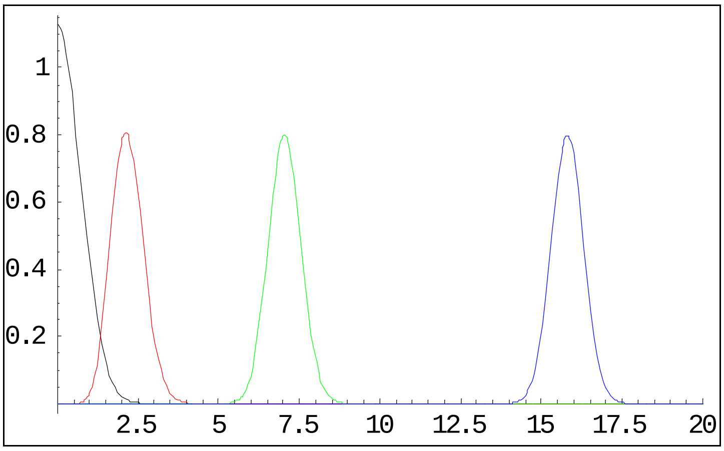


Figure A.2: Distribution of the length in N dimensions (curves from left to right for N = 1, 10, 100 and 500); abscissa: multiple of σ)

From Figure A.2 it is apparent that when one is dealing with a high dimensional Gaussian distribution numerically and wants to include the dominant portion of the probability distribution of length or RMS into a certain calculation, the conventional 99% cutoff at 3 σ for single dimension must be extended. This extension can be drastic with large N. It is on the other hand awkward to solve for the cutoff for a given value of accumulated distribution by integrating (I.13) and inverting the resulting gamma function. An efficient way to obtain such cut-offs, with N up to a few hundred, is by repeatedly differentiating (I.13) for higher order inflection points, which include progressively more of the distribution and the process converges rapidly within the first few steps. Table A.1 illustrates this technique on an RMS distribution over 500-dimensions. More than 98 % of the distribution is sandwiched between the two 4th order inflection points, as is evident from both Figure A.2 and Table A.1. The computation takes less than 10 seconds[27]. This technique was used in determining the magnitude of the overall input orbit RMS in the main analysis.

Order of Derivative	1	2	3	4
Lower Bound Cutoff (σ)	15.795569	15.295504	14.932080	14.634732
Integral below Low Cutoff	0.495789	0.154742	0.039693	0.009072
Upper Bound Cutoff (σ)	15.795569	16.295630	16.664324	16.969316
Integral below High Cutoff	0.495789	0.837593	0.956548	0.989548

Table A.1 Inclusion of a 500-dimensional RMS distribution based on 4th order derivative cut-offs.

[27] This is timed on a Pentium III 600 MHz PC running *Mathematica 4.0*. The next order includes > 99.5% of the distribution after 12.5 seconds of computation.

J Exception Handling due to Rank Deficiency

In all algorithms discussed above one occasionally encounters degenerate or near- degenerate configurations due to the large dimensionality involved in the calculations. Many of the inversion operations fail when this happens and special attention must be paid to handle such cases. Two cases are of particular interest:

Rank deficiency:

It is granted that in an over-constraint projection the mapping matrix takes the actuator into a subspace of the responder space, and the algorithms developed above can evaluate this subspace. On the other hand, due to rank deficiency in the map, sometimes even critically or under-constrained maps can project into a subspace of the responder space, and matrix inversions fail in the absence of special exception handling. This problem has been addressed in this analysis by monitoring the true dimensionality of the image subspace of all critically or under-constrained maps. If this is found to be less than the rank of the matrix, a different algorithm is used which works in the true image subspace of a lower dimensionality.

Near-singularity:

This happens especially with large numbers of elements making singular combinations more likely. It does not imply rank deficiency but can cause the same problem when the finite numerical accuracy cannot maintain linear independence. The algorithm used in the current analysis monitors a potential near-degeneracy of such projections using SVD, and eliminates the singularity [7] to pre-empt such problems.