Closed-orbit correction in synchrotrons*

D. Dinev
Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, Sofia

Fiz. Élem. Chastits At. Yadra **28**, 1013–1060 (July–August 1997)

Algorithms for closed-orbit correction in synchrotrons and related topics such as error sources, statistical characteristics of the orbit, etc., are discussed. The review covers both traditional methods for orbit correction (beam-bump, harmonic correction, etc.) and new developments (eigenvector correction, SVD, algorithms for optimum positioning of dipoles, etc.). The use of expert systems and artificial neural networks is described as well. The last section is devoted to first-turn steering. © *1997 American Institute of Physics.* [S1063-7796(97)00404-X]

1. INTRODUCTION

The transverse motion of the particles in a cyclic accelerator is a superposition of motion along a closed trajectory (equilibrium or closed orbit) and betatron and radial-phase oscillations around this curve. In a real magnetic structure both the closed orbit and the oscillations around it are distorted by errors in the magnetic fields, displacements and tilts of the elements from their design positions, stray fields, and ground movements. As far as the closed orbit is concerned, perturbations cause its deformation. The maximum deviation of the distorted orbit from the reference orbit reaches some ten millimeters. Such a large deviation does not allow the accelerator aperture to be used effectively and hampers the work of the injection and extraction systems.

In storage rings the beam lifetime and the maximum current of the accumulated particles depend on the accuracy of the orbit.

An important point is the orbit stability. The time changes of the orbit increase the dynamic aperture. In synchrotron light sources an unstable orbit will increase the effective emittance, thus reducing the effective brightness of the photon beams.

This is the reason why in accelerators special systems of either small correcting magnets or additional correcting coils in the main dipoles or controlled displacements of the quadrupoles are used.

The purpose of the orbit correction consists in choosing the proper strengths and positions of the correction elements so that the smallest possible deviations from the reference orbit may be achieved.

A special problem is first-turn steering, which we discuss in the last section.

In this paper we give a survey of both the major orbit-correction algorithms used in synchrotrons and storage rings and original results on the orbit correction obtained by the author during his work on the superconducting synchrotron Nuclotron at JINR (Dubna) and on the cooler synchrotron COSY (Jülich).

2. TRANSVERSE PARTICLE MOTION UNDER LINEAR PERTURBATIONS

We shall begin our discussion of the closed-orbit distortion and correction with a brief survey of the transverse particle motion in cyclic accelerators in the presence of so-called linear perturbations.

In the curvilinear coordinate system (s,x,y) used in cyclic accelerators, where s is directed along the reference orbit, x along the main normal vector, and z along to the binormal vector, the Lagrangian of the transverse particle motion can be represented in the form[1]

$$\mathscr{L}(x,x',z,z',s)=p\sqrt{\left(1+\frac{x}{\rho}\right)^2+x'^2+z'^2}+e\left[\left(1+\frac{x}{\rho}\right)A_s+x'A_x+z'A_z\right]. \tag{2.1}$$

In (2.1) the distance along the reference orbit s was taken to be the independent variable; the primes denote differentiation with respect to s; p and e are the particle momentum and charge; ρ is the radius of curvature of the reference orbit; and A_s, A_x, A_z are the vector-potential components.

From (2.1) the following expression for the transverse-motion Hamiltonian can be deduced:

$$\mathscr{H}(x,p_x,z,p_z,s)=-\left(1+\frac{x}{\rho}\right)\times\sqrt{p^2-(p_x-eA_x)^2-(p_z-eA_z)^2}-e\left(1+\frac{x}{\rho}\right)A_s. \tag{2.2}$$

We shall consider here the cases of a sector magnetic field and of a quadrupole field, which are the most important ones for particle accelerators. For these cases, by substituting in (2.2) the relevant expressions for the vector potential A and expanding the kinematic part in a power series, retaining only the major first few terms, one finds

$$\mathscr{H}\approx\mathscr{H}_0=\begin{cases}\dfrac{1}{2}\dfrac{p_x^2}{p_0}+\dfrac{1}{2}\dfrac{p_z^2}{p_0}+p_0\dfrac{1-n}{2}\left(\dfrac{x}{\rho}\right)^2+p_0\dfrac{n}{2}\left(\dfrac{z}{\rho}\right)^2 \\ \quad \text{for a sector magnetic field,} \\ \dfrac{1}{2}\dfrac{p_x^2}{p_0}+\dfrac{1}{2}\dfrac{p_z^2}{p_0}-\dfrac{eg}{2}(x^2-z^2) \\ \quad \text{for a quadrupole field.}\end{cases} \tag{2.3}$$

In (2.3) we denoted by $\mathscr{H}_0$ the Hamiltonian of the linearized transverse motion; $p_0=-eB_0\rho$ is the particle mo-

 1063-7796/97/040398-20$10.00

mentum corresponding to the reference trajectory; $B_0\rho$ is the beam rigidity; $n=-(\rho/B_0)\,(\partial B/\partial x)$ is the field index; and g is the quadrupole gradient.

We shall consider in this paper correcting magnets as short magnets, with a small ($B_c \ll B_0$) and uniform ($n=0$) field. Thus the linearized Hamiltonian for the corrections is

$$\mathcal{H}_0=\frac{1}{2}\frac{p_x^2}{p_0}+\frac{1}{2}\frac{p_z^2}{p_0}-eB_c x. \tag{2.4}$$

Let us now introduce in the Hamiltonians the so-called linear perturbations:

a) field errors, $\Delta B=B-B_q$;

b) magnetic-element misalignments, Δx and Δz;

c) dipole tilts around the axis s, θ;

d) stray magnetic field, ΔB.

These errors cause terms linear in x and z to appear in the Hamiltonians:

$$\mathcal{H}=\mathcal{H}_0+\delta\mathcal{H}^{(1)}, \tag{2.5}$$

where

$$\delta\mathcal{H}^{(1)}=\begin{cases} -e\Delta B_z x+\dfrac{p_0}{\rho}\,\theta z+\dfrac{p_0(1+n)}{\rho^2}\,x\Delta x \\ \quad -\dfrac{p_0 n}{\rho^2}\,z\Delta z+\left(1+\dfrac{x}{\rho}\right)\Delta p \quad \text{for dipoles,} \\ egx\Delta x-egz\Delta z \quad \text{for quadrupoles.} \end{cases} \tag{2.6}$$

The quadratic part $\mathcal{H}_0$ of the Hamiltonian (2.5) describes particle oscillations, as long as the linear part $\delta\mathcal{H}^{(1)}$ causes an internal force depending on the variable s. This is the well-known classical-mechanics case of forced small oscillations. Because of the internal force, the orbit is distorted. The new orbit is a periodic solution of Hamilton's equations:

$$\frac{dx_{c0}}{ds}=\frac{\partial\mathcal{H}}{\partial p_{x,c0}},$$

$$\frac{dp_{x,c0}}{ds}=-\frac{\partial\mathcal{H}}{\partial x_{c0}},$$

$$x_{c0}(s+2\pi R)=x_{c0}(s),$$

$$p_{x,c0}(s+2\pi R)=p_{x,c0}(s). \tag{2.7}$$

In (2.7), R is the mean radius of the accelerator.

Equations equivalent to (2.7) are valid for the vertical plane as well.

Let u and p_u be the conjugate variables describing the oscillations around the orbit:

$$x=x_{c0}+u,$$

$$p=p_{x,c0}+p_u. \tag{2.8}$$

We now perform a canonical transformation from the variables x, p_x to u, p_u, using as a generating function

$$f_2(x,p_u,s)=(p_{x,c0}(s)+p_u)x-x_{c0}(s)p_u. \tag{2.9}$$

We obtain for the new Hamiltonian

$$K(u,p_u,v,p_v,s)=\mathcal{H}(x_{c0}(s)+u,p_{x,c0}(s)+p_u,z_{c0}(s)+v,p_{z,c0}(s)+p_v,s)$$
$$=\mathcal{H}-\left(\frac{\partial\mathcal{H}}{\partial x_{c0}}\,u+\frac{\partial\mathcal{H}}{\partial p_{x,c0}}\,p_u+\frac{\partial\mathcal{H}}{\partial z_{c0}}\,v+\frac{\partial\mathcal{H}}{\partial p_{z,c0}}\,p_v\right). \tag{2.10}$$

The Hamiltonian K does not contain terms linear in u, p_u, v, p_v, i.e., it describes only particle oscillations around the closed orbit.

From the Hamiltonians (2.3) and (2.6) one can deduce the equations of the transverse particle motion. In Newton's form they are

$$\frac{d^2x}{ds^2}+k_x(s)x=F_x(s),$$

$$\frac{d^2z}{ds^2}+k_z(s)z=F_z(s), \tag{2.11}$$

where

$$k_x(s)=\begin{cases} \dfrac{(1-n)}{\rho^2} \text{ for dipoles,} \\ 0 \text{ for correctors,} \\ -\dfrac{g}{B_0\rho} \text{ for quadrupoles,} \end{cases} \tag{2.12}$$

$$k_z(s)=\begin{cases} \dfrac{n}{\rho^2} \text{ for dipoles,} \\ 0 \text{ for correctors,} \\ \dfrac{g}{B_0\rho} \text{ for quadrupoles,} \end{cases} \tag{2.13}$$

$$F_x(s)=\begin{cases} -\dfrac{\Delta B_z}{B_0\rho}-\dfrac{(1+n)}{\rho^2}\,\Delta x \text{ for dipoles,} \\ -\dfrac{B_c}{B_0\rho} \text{ for correctors,} \\ -\dfrac{g\Delta x}{B_0\rho} \text{ for quadrupoles,} \end{cases} \tag{2.14}$$

$$F_z(s)=\begin{cases} -\dfrac{n}{\rho^2}\,\Delta z-\dfrac{\theta}{\rho} \text{ for dipoles,} \\ -\dfrac{B_c}{B_0\rho} \text{ for correctors,} \\ -\dfrac{g\Delta x}{B_0\rho} \text{ for quadrupoles.} \end{cases} \tag{2.15}$$

3. CLOSED ORBIT

The equations in (2.11) are Hill's equation with a nonzero right-hand side. Its general solution can be represented as a sum of the general solution of the homogeneous equation and a particular solution of the nonhomogeneous one. The latter can be taken as periodic, bearing in mind the accelerator symmetry. This periodic particular solution will describe a closed orbit as long as the general solution of the homogeneous equation describes the particle oscillation.

The particle oscillations in an accelerator can be described by Twiss's amplitude function $\beta(s)$:[2]

$$x(s)=a\sqrt{\beta(s)}\cos\left(Q\int_0^s\frac{ds}{Q\beta(s)}\right), \tag{3.1}$$

where Q is the number of the betatron oscillations per turn.

Let us introduce the following new variables:

a) the generalized azimuth

$$\varphi=\int_0^s\frac{ds}{Q\beta(s)}; \tag{3.2}$$

b) the normalized deviation

$$\eta=\frac{x}{\sqrt{\beta(s)}}. \tag{3.3}$$

In these new variables the oscillations (3.1) can be written as

$$\eta(\varphi)=a\cos(Q\varphi). \tag{3.4}$$

By implication we shall describe the closed orbit using the variables (3.2) and (3.3).

Using the properties of the β function, it is possible to transform (2.11) to an equation of forced oscillations:

$$\frac{d^2\eta}{d\varphi^2}+Q^2\eta=Q^2f(\varphi), \tag{3.5}$$

where

$$f(\varphi)=\beta^{3/2}(\varphi)F(\varphi). \tag{3.6}$$

Using the method of varying the integration constants, the following periodic particular solution of (3.5) can be obtained:[2]

$$\eta=\frac{Q}{2\sin\pi Q}\int_\varphi^{\varphi+2\pi}f(t)\cos Q(\varphi+\pi-t)dt. \tag{3.7}$$

The integral representation (3.7) is the basic formula for the description of the closed orbit.

There exist two approaches to the treatment of the closed orbit:

a) *A matrix approach.*

As the perturbations are equal to zero outside the magnetic elements $[f(\varphi)=0]$ and the elements are sufficiently short compared with the accelerator circumference ($\Delta\varphi \ll 2\pi$), the integral (3.7) can be transformed into the sum

$$\eta(\varphi_i)=\frac{Q}{2\sin\pi Q}\sum_{\substack{j=1\\ \varphi_i\leq\varphi_j\leq\varphi_i+2\pi}}^{M+L}\overline{f}_i\Delta\varphi_j\cos Q(\varphi_i+\pi-\varphi_j). \tag{3.8}$$

Here M is the total number of dipoles, L is the number of quadrupoles, and a bar above a symbol denotes averaging over the magnetic element.

It is convenient to put (3.8) in the matrix form

$$\eta_i=\sum_{\substack{j=i\\ \varphi_i\leq\varphi_j\leq\varphi_i+2\pi}}^{M+L}A_{ij}\delta_j, \tag{3.9}$$

where

$$A_{ij}=\cos Q(\varphi_i+\pi-\varphi_j) \tag{3.10}$$

and

$$\delta_j=\begin{cases}-\dfrac{Q\overline{\beta_j^{3/2}}\Delta\varphi_j}{2\sin\pi Q}\Delta B_j-\dfrac{Q\overline{\beta_j^{3/2}}\eta_j\Delta\varphi_j}{2\sin\pi Q\rho_j^2}\Delta x_j & \text{for dipoles,}\\ -\dfrac{Q\overline{\beta_j^{3/2}}\Delta\varphi_j}{2\sin\pi Q}B_{c,j} & \text{for correctors,}\\ -\dfrac{Q\overline{\beta_j^{3/2}}g_j\Delta\varphi_j}{2\sin\pi QB\rho}\Delta x_j & \text{for quadrupoles,}\end{cases} \tag{3.11}$$

$$\delta=\frac{\sqrt{\beta}}{2\sin\pi Q}\varepsilon, \tag{3.12}$$

in which ε is the kick in the element.

The reason for introducing the generalized perturbations δ_j is that they have the same dimensions as the normalized orbit η, i.e., $m^{1/2}$.

b) *Harmonic-analysis approach.* The orbit $\eta(\varphi)$ is periodic with period 2π. Let us expand it in a Fourier series:

$$\eta(\varphi)=\frac{u_0}{2}+\sum_{k=1}^{\infty}(u_k\cos k\varphi+v_k\sin k\varphi). \tag{3.13}$$

Let us also expand the perturbations $f(\varphi)$ in a Fourier series:

$$f(\varphi)=\frac{f_0}{2}+\sum_{k=1}^{\infty}(f_k\cos k\varphi+g_k\sin k\varphi). \tag{3.14}$$

From Eq. (3.5) the following relations between the Fourier coefficients of the orbit and of the perturbations can be deduced:

$$u_0=f_0,$$
$$u_k=\left(\frac{Q^2}{Q^2-k^2}\right)f_k,$$
$$v_k=\left(\frac{Q^2}{Q^2-k^2}\right)g_k. \tag{3.15}$$

Of course, the matrix and the harmonic-orbit treatments give the same results, as can be demonstrated by using the formula

$$\cos Q(\varphi_i+\pi-\varphi_j)=(\sin\pi Q)\frac{2Q}{\pi}\left(\frac{1}{2Q^2}-\sum_{k=1}^{\infty}\cos k\frac{(\varphi_i-\varphi_j)}{k^2-Q^2}\right). \tag{3.16}$$

4. ERROR SOURCES

The linear perturbations causing distortion of the closed orbit can be summarized in the following way.

a) *Constant errors:*

i) errors in the coercive force, which can be estimated approximately as

$$\Delta B\approx-\mu_0\frac{l_{st}}{l_a}\Delta H_c, \tag{4.1}$$

where l_{st} is the magnet core length and l_a is the aperture;

ii) errors due to eddy currents;

iii) stray magnetic fields;

iv) earth magnetic field.

b) *Errors proportional to the main magnetic field:*

i) permeability errors, given approximately by

$$\frac{\Delta B}{B} \approx \left(\frac{l_{st}}{\mu_r^2 l_a}\right) \Delta\mu; \tag{4.2}$$

ii) errors in the magnet core length,

$$\frac{\Delta B}{B} \approx -\frac{1}{\mu_r l_a + l_{st}} \Delta l_{st}; \tag{4.3}$$

iii) aperture errors,

$$\frac{\Delta B}{B} \approx -\frac{1}{\frac{l_{st}}{\mu_r} + l_a} \Delta l_a; \tag{4.4}$$

iv) adjustment errors—element misalignments, median-plane displacements, dipole tilts, quadrupole magnetic-center displacements, errors in the coil positions, etc.;

v) ground movement.

c) *Errors appearing only in high fields.* These are errors due to saturation.

d) *Fast noise.* These are ground movements and power-supply ripples in the 1–100 Hz bandwidth which cause fluctuations of the orbit.

5. STATISTICAL CHARACTERISTICS OF THE PERTURBATIONS AND THE ORBIT

The perturbations ΔB and Δx are random functions of the generalized azimuth φ. Bearing in mind that they are nonzero only within the magnetic elements and that these elements are sufficiently short compared with the accelerator circumference ($\Delta\varphi \ll 2\pi$), one can represent the perturbations as sums of elementary random functions: $\Delta B_i \Pi_i(\varphi)$ and $\Delta x_i \Pi_i(\varphi)$, where ΔB_i and Δx_i are random variables and $\Pi_i(\varphi)$ are single rectangular pulses of length $\Delta\varphi_i$. The random variables ΔB_i and Δx_i are uncorrelated and normally distributed, with zero mean and identical standard deviations $\sigma_{\Delta B}$, $\sigma_{\Delta x}$ for all elements.

Let us expand the perturbations in a Fourier series:

$$\Delta B(\varphi) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos k\varphi + b_k \sin k\varphi). \tag{5.1}$$

The Fourier amplitudes a_k and b_k are random variables. Their means vanish, and their variances are

$$D(a_0) = \sigma_{\Delta B}^2 \sum_{j=1}^{M} \left(\frac{\Delta\varphi_j}{2\pi}\right)^2,$$

$$D(a_k) = \sigma_{\Delta B}^2 \sum_{j=1}^{M} \left(\frac{\cos k\varphi_j \Delta\varphi_j}{\pi}\right)^2,$$

$$D(b_k) = \sigma_{\Delta B}^2 \sum_{j=1}^{M} \left(\frac{\sin k\varphi_j \Delta\varphi_j}{\pi}\right)^2. \tag{5.2}$$

In uniform magnetic structures consisting of p equal periods the formulas (5.2) are simplified, leading to a "white" spectrum:

$$D(a_k) = D(b_k) = \frac{1}{2} D(a_0) = \frac{\sigma_{\Delta B}^2 P}{2\pi^2} \sum_{j=1}^{M_p} \Delta\varphi_j^2. \tag{5.3}$$

In (5.3), M_p is the number of dipoles per period. Using the accelerator symmetry, it is also possible to demonstrate that $E(a_k, b_k) = 0$, i.e. that a_k and b_k are uncorrelated. Analogous results can be obtained for Δx.

Let us now discuss the statistical characteristics of the orbit itself.

From (3.9) it follows that the orbit deviation η_i as a sum of normally distributed values is normally distributed itself, with mean equal to zero and variance

$$D(\eta_i) = \left(\frac{Q}{2B_0\rho \sin \pi Q}\right)^2 \left[\sigma_{\Delta B}^2 \sum_{j=1}^{M} \beta_j^3 \Delta\varphi_j^2 \cos^2 Q(\varphi_i + \pi - \varphi_j) + \sigma_{\Delta x}^2 \sum_{j=1}^{L} g_j^2 \beta_j^3 \Delta\varphi_j^2 \cos^2 Q(\varphi_i + \pi - \varphi_j)\right]. \tag{5.4}$$

In uniform magnetic structures, (5.4) is simplified:

$$D(\eta_i) = \left(\frac{Q}{2B_0\rho \sin \pi Q}\right)^2 p\sigma_{\Delta B}^2 \sum_{j=1}^{M_p} \frac{\beta_j^3 \Delta\varphi_j^2}{2} + \left(\frac{Q}{2B_0\rho \sin \pi Q}\right)^2 p\sigma_{\Delta x}^2 \sum_{j=1}^{L_p} \frac{g_j^2 \beta_j^3 \Delta\varphi_j^2}{2}. \tag{5.5}$$

In the same way, the statistical characteristics of the orbit divergence $\eta'(\varphi)$ can be calculated, and one obtains, for uniform structures,

$$D(\eta') = Q^2 D(\eta). \tag{5.6}$$

Successful accelerator operation requires that the maximum orbit deviation $\eta_{\max}$ and its statistical characteristics be known.

As for every accelerator from a large group of accelerators with different random error distributions, the maximum orbit deviation $\eta_{\max}$ appears at a different point $\varphi_{\max}$. The quantity $\eta_{\max}$ is not normally distributed.

It is very difficult to obtain the statistical characteristics of $\eta_{\max}$ in the general case. As a rough approximation, let us restrict ourselves to the major $k \approx Q$ harmonic in the orbit expansion (3.12):

$$\eta(\varphi) \approx u_k \cos k\varphi + v_k \sin k\varphi = A \cos(Q\varphi + \alpha_k). \tag{5.7}$$

It is easy to see that in this approximation $E(\eta, \eta') = 0$, i.e., η and η' are statistically independent.

For the two-dimensional probability we have[3]

$$p\left(\eta, \frac{\eta'}{Q}\right) = \frac{1}{2\pi\sigma_\eta^2} \exp\left(-\frac{A^2}{2\sigma_\eta^2}\right). \tag{5.8}$$

Obviously, the orbit amplitude probability $p(A)$ can be obtained by integrating (5.8) along a circle of radius A in the $(\eta, \eta'/Q)$ plane:

$$p(A)=\int_0^{2\pi} p\left(\eta, \frac{\eta'}{Q}\right)A\,d\varphi=\frac{A}{\sigma_\eta^2}\exp\left(-\frac{A^2}{2\sigma_\eta^2}\right). \quad (5.9)$$

Equation (5.9) represents a Rayleigh probability distribution

$$p(A)=\frac{2A}{\sigma_A^2}\exp\left(-\frac{A^2}{\sigma_A^2}\right) \quad (5.10)$$

with

$$\sigma_A=\sqrt{2}\sigma_\eta=\sqrt{2}\sigma_{u_k}=\sqrt{2}\sigma_{v_k}. \quad (5.11)$$

The cumulative distribution function for the Rayleigh distribution is

$$\phi(A)=1-\exp\left(-\frac{A^2}{\sigma_A^2}\right). \quad (5.12)$$

A better estimate is obtained if two contributing harmonics $k<Q<(k+1)$ are taken into account.

In this approximation,

$$\begin{aligned}\eta(\varphi)\approx{}&u_k\cos k\varphi+v_k\sin k\varphi+u_{k+1}\cos(k+1)\varphi\\&+v_{k+1}\sin(k+1)\varphi=(u_k+u_{k+1}\cos\varphi\\&+v_{k+1}\sin\varphi)\cos k\varphi+(v_k-u_{k+1}\sin\varphi\\&+v_{k+1}\cos\varphi)\sin k\varphi,\end{aligned} \quad (5.13)$$

i.e., we have a harmonic oscillation with a slowly changing amplitude $A(\varphi)$:

$$\begin{aligned}A^2(\varphi)=(u_k+u_{k+1}\cos\varphi+v_{k+1}\sin\varphi)^2+(v_k\\-u_{k+1}\sin\varphi+v_{k+1}\cos\varphi)^2.\end{aligned} \quad (5.14)$$

From (5.14) the maximum amplitude can be calculated:

$$A=\max A(\varphi)=r_k+r_{k+1}, \quad (5.15)$$

where

$$r_k^2=u_k^2+v_k^2;\quad r_{k+1}^2=u_{k+1}^2+v_{k+1}^2. \quad (5.16)$$

As has already been demonstrated above, τ_k and τ_{k+1}, which are the random amplitudes of the kth and $(u+1)$th harmonics, have Rayleigh distributions. These amplitudes are statistically independent. From these assumptions the following expression for the variance of the maximum amplitude can be derived:

$$\sigma_A^2=R_k^2+\frac{\pi}{2}R_kR_{k+1}+R_{k+1}^2, \quad (5.17)$$

where

$$R_k^2=2\sigma_{u_k}^2;\quad R_{k+1}^2=2\sigma_{u_{k+1}}^2. \quad (5.18)$$

In Ref. 4 it was also shown that the probability for the orbit amplitude to be greater than A is

$$\begin{aligned}F(A)=2\sqrt{\pi}\,\frac{R_kR_{k+1}}{R_k^2-R_{k+1}^2}\,\frac{A}{\sqrt{R_k^2+R_{k+1}^2}}\\\times\exp\left(-\frac{A}{R_k^2+R_{k+1}^2}\right).\end{aligned} \quad (5.19)$$

It was shown in Ref. 4 that the best agreement between the analytical estimates and the results of the computer simulation of the closed orbit can be achieved if the major three or four harmonics of the perturbations are taken into account. As the exact solution in this approximation is accompanied by great mathematical difficulties, only an approximate estimation for the orbit-amplitude variance is carried out:

$$\frac{\sigma_A^2}{\Sigma R_j^2}=\left(1+\frac{\pi}{2}\frac{\Sigma_{i,j}^{i\neq j}R_iR_j}{\Sigma R_j^2}\right), \quad (5.20)$$

$$\begin{aligned}F(A)=&(4\pi)^{(m-1)/2}\\&\times\prod_{j=1}^{m}\left(\frac{R_j}{\sqrt{\Sigma R_j^2}}\right)\exp\left(-\frac{A^2}{\Sigma R_j^2}\right)\\&\times\sum_{i=0}^{(m-1)/2}\frac{(-1)^i(m-1)!}{2^{2i}i!(m-1-2i)!}\left(\frac{A}{\sqrt{\Sigma R_j^2}}\right)^{m-1-2j},\end{aligned} \quad (5.21)$$

where m is the number of harmonics taken into account.

6. GENERAL DESCRIPTION OF THE ORBIT CORRECTION METHODS

As a rule, in accelerators the number of beam position monitors (N) is less than the numbers of dipoles (M) and quadrupoles (L): $N<M+L$. This means that from the readings of the BPMs we can calculate only a part of the perturbations—correction with uncertainty.

A general block diagram of the orbit correction is shown in Fig. 1. Figure 2 gives a classification of the orbit correction methods.

The orbit correction methods can be divided into two main groups: methods for local correction and methods for correction of the orbit over the whole ring (general correction).

The local correction methods correct the orbit only in a part of the accelerator circumference. Outside this region, the orbit is unchanged.

The methods for general correction cover methods that try to compensate perturbations directly and methods in which a kind of goal function, depending on the orbit deviations and corrector strengths, is minimized.

The methods with perturbation compensation try to assess the perturbation strengths.

In the harmonic correction, this is achieved through the approximate values of the first perturbation harmonics. In the method proposed by G. Guignard and its later improvements, the "probability" of having errors in a given area of the accelerator circumference is found. Finally, in Warren and Channell's suggestion, the values of all the errors are calculated by applying a special measurement procedure.

After the perturbations have been assessed, corrections with proper values and positions are applied in order to compensate the orbit deviations.

In the methods with orbit deviation compensation a quality criterion is formulated. It can be either a function or a functional of the orbit deviations and corrector strengths. Then the minimum of the quality criterion is chosen. Single-step, multistep, and dynamical methods are included in this group.

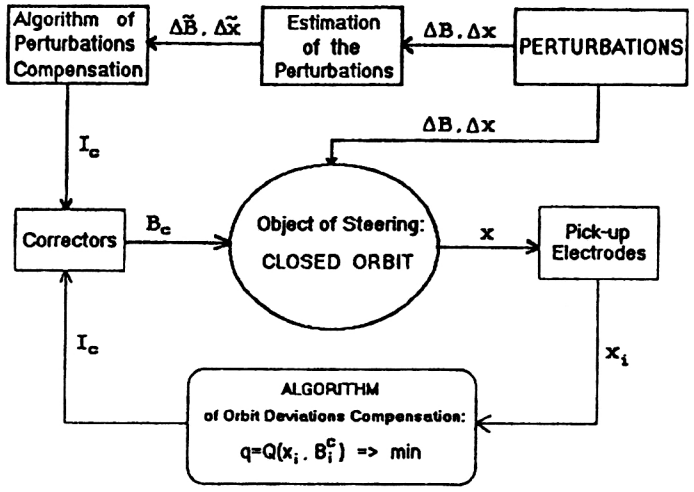

FIG. 1. Block diagram of the orbit correction.

In the single-step method, the orbit is considered only at the points where BPMs are situated and is characterized by the state vector $\mathbf{x}=(x_1,x_2,\ldots,x_n)$, where the x_i are the orbit deviations. A goal function

$$q=Q(\mathbf{x},\mathbf{B}_c) \tag{6.1}$$

is formed.

In (6.1), $\mathbf{B}_c$ is a vector whose components are corrector strengths. They are determined so that q has a minimum.

In the multistep methods, the whole accelerator circumference is divided into separate areas. Initially, the orbit over the first area is corrected. Then we go to the next area, taking into consideration the results from the correction in the previous step, and so on.

The dynamical methods of correction treat the orbit as a whole curve $\eta(\varphi)$ rather than just orbit deviations in the BPMs, $\eta_i=\eta(\varphi_i)$. The quality criterion is a functional:

$$I=\int_0^{2\pi} Q(\eta(\varphi),B_c(\varphi))d\varphi. \tag{6.2}$$

The condition for I to have a minimum gives us the chosen corrector strengths.

7. CORRECTION METHODS WITH COMPENSATION OF THE PERTURBATIONS

7.1. Harmonic correction method

In the harmonic correction method, the orbit Fourier spectrum

$$\eta(\varphi)=\frac{u_0}{2}+\sum_{k=1}^{\infty}(u_k\cos k\varphi+v_k\sin k\varphi) \tag{7.1.1}$$

is first calculated.

As we know, the methods of applied harmonic analysis must be used only for the orbit deviations at the positions of the BPMs and not for the whole curve $\eta(\varphi)$. By measuring the orbit with $2N$ BPMs we can determine only approximate values of the first N orbit Fourier harmonics. In the case of uniformly situated monitors the simplest method is to approximate the first orbit harmonics with the so-called Bessel coefficients:[5]

$$u_0\approx U_0=\frac{1}{N}\sum_{i=1}^{2N}\eta_i,$$

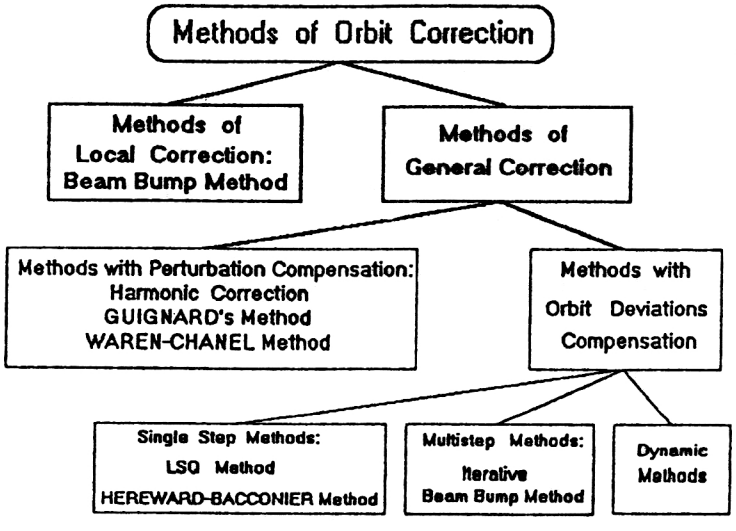

FIG. 2. Classification of orbit correction methods.

$$u_k \approx U_k = \frac{1}{N} \sum_{i=1}^{2N} \eta_i \cos k\varphi_i \quad (k=1,2,\ldots,N),$$

$$v_k \approx V_k = \frac{1}{N} \sum_{i=1}^{2N} \eta_i \sin k\varphi_i \quad (k=1,2,\ldots,N-1). \quad (7.1.2)$$

In the general case of nonequidistant monitors the trapezoidal rule for integral approximations gives

$$u_k \approx \frac{1}{\pi} \sum_{i=1}^{N} \eta_i \cos k\phi_i \left(\frac{\phi_{i+1} - \phi_{i-1}}{2} \right),$$

$$v_k \approx \frac{1}{\pi} \sum_{i=1}^{N} \eta_i \sin k\phi_i \left(\frac{\phi_{i+1} - \phi_{i-1}}{2} \right),$$

$$\phi_0 = \phi_N - 2\pi, \quad \phi_{N+1} = \phi_1 + 2\pi. \quad (7.1.3)$$

In this general case the LSQ criterion of approximation

$$\sum_{i=1}^{p} \left[\eta_i - \sum_{k=0}^{n} c_k \psi_k(\phi_i) \right]^2 \to \min, \quad (7.1.4)$$

where the c_k are Fourier coefficients and ψ_k is the system of trigonometric functions, gives

$$\begin{bmatrix} u_0 \\ u_1 \\ \cdot \\ \cdot \\ u_N \\ v_1 \\ \cdot \\ \cdot \\ v_{N-1} \end{bmatrix} = (S^T S)^{-1} S^T \begin{bmatrix} \eta_1 \\ \eta_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \eta_{2N} \end{bmatrix}, \quad (7.1.5)$$

where

$$S = \begin{bmatrix} 1/2 & \cos\phi_1 & \cos 2\phi_1 & \cdots & \cos N\phi_1 & \sin\phi_1 & \cdots & \sin(N-1)\phi_1 \\ 1/2 & \cos\phi_2 & \cos 2\phi_2 & \cdots & \cos N\phi_2 & \sin\phi_2 & \cdots & \sin(N-1)\phi_2 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot \\ 1/2 & \cos\phi_{2N} & \cos 2\phi_{2N} & \cdots & \cos N\phi_{2N} & \sin\phi_{2N} & \cdots & \sin(N-1)\phi_{2N} \end{bmatrix}. \quad (7.1.6)$$

Knowing the orbit harmonics, the perturbation harmonics can be calculated by using (3.14).

In the harmonic correction method, the fields at the $2N$ correcting dipoles are chosen so that the first N perturbation harmonics are canceled.[6-8]

Bearing in mind that the correcting dipoles are short ($\Delta\varphi_c \ll 2\pi$), and transforming the integrals to sums, we can then solve the following system of equations:

$$\frac{1}{\pi} \sum_{i=1}^{2N} f_i^c \Delta\varphi_i^c = U_0,$$

$$\frac{1}{\pi} \sum_{i=1}^{2N} f_i^c \cos k\varphi_i \Delta\varphi_i = \frac{Q^2 - k^2}{Q^2} U_k,$$

$$\frac{1}{\pi} \sum_{i=1}^{2N} f_i^c \sin k\varphi_i \Delta\varphi_i = \frac{Q^2 - k^2}{Q^2} V_k. \quad (7.1.7)$$

Then the fields at the correcting dipoles, B_i^c ($i = 1,2,\ldots,2N$), can be calculated from the generalized perturbations f_i^c by using (2.14), (2.15), and (3.6).

7.2. Guignard's method

It is very important that the largest error sources be found. Then we can check more carefully the corresponding elements, and experience shows that in many cases, after removing this very rough imperfection, the maximum orbit deviation is reduced to sufficiently small values. As in an error-free area of the accelerator circumference, the right-hand side of Eq. (3.5) is equal to zero. The orbit in this error-free area is

$$\eta = A \cos Q\varphi + B \sin Q\varphi + C. \quad (7.2.1)$$

The orbit (7.2.1) must be matched to the whole orbit at both ends of the area under consideration. Then the constants A,B,C are calculated.

If now we assume that there are errors in the above area, then the smooth solution (7.2.1) will no longer be valid. At the points with errors, η' will undergo kicks.

Obviously, the divergence of the real measured orbit from the smooth model (8.2.1) will be a measure of the perturbation strengths in the area.

Following Guignard,[9] we define the "probability" for the existence of perturbations in an area of the accelerator as

$$\Psi = \frac{1}{n} \sum_{k=p}^{p+n-1} [\sqrt{\beta_k}(A \cos Q\varphi_k + B \sin Q\varphi_k + C) - x_k]^2. \quad (7.2.2)$$

In (7.2.2), p is the number of the first BPM and $n \geq 4$ is the total number of BPMs included in the area.

The constants A,B,C in (7.2.2) define the unperturbed orbit. We shall determine their values so that the unperturbed orbit will lie as close as possible to the readings x_k in the BPMs (in the LSQ sense). In other words, we calculate A, B, C from the condition $\Psi \rightarrow \min$.

In Guignard's method, we calculate the "probability" Ψ for successive areas of the accelerator ring. The areas with large Ψ values are possible sources of errors.

7.3. Warren and Channell's suggestion

As has already been mentioned, in general the number of BPMs (N) is less than the numbers of dipoles (M) and quadrupoles (L): $N<M+L$. Therefore we are not able to solve the system of equations (3.9) for the errors. In Ref. 10 Warren and Channell suggest enlarging this system by adding some new equations. For this purpose, new measurements of the orbit have to be made by changing the synchrotron parameters and keeping the errors unchanged. A new set of orbit measurements can be carried out by changing the quadrupole strengths (i.e., changing Q) or changing the signs of the quadrupole gradients (for example, reducing FODO structure into DOFO). The enlarged system of equations is solved by the LSQ method.

8. LOCAL CORRECTION METHODS

8.1. Beam-bump method

At some positions (injection, extraction, and others) a very high degree of orbit correction is necessary. On the other hand, at other parts of the accelerator ring the orbit deviation may be much greater than the average one, owing to strong local stray fields, strong local imperfections, or ground movement. These require the development of methods for local correction.

One such local correction method is the beam-bump method, suggested by Collins in the sixties.[11-14]

The idea of the beam-bump method consists in a local orbit correction by means of three correcting dipoles (Fig. 3). The orbit deviation is compensated in a BPM or pick-up electrodes (PUE) situated near the middle corrector, as long as outside the area occupied by the correctors the orbit is kept unchanged. Therefore we have the conditions

$$x(\varphi_{\mathrm{PUE}}) = -x_m,$$
$$x(\varphi<\varphi_1, \varphi>\varphi_3) = x'(\varphi<\varphi_1, \varphi>\varphi_3) = 0. \tag{8.1.1}$$

From (8.1.1) and (3.9) the following system of three equations for the corrector strengths can be deduced:

$$\sin\mu_{12}\sqrt{\beta_2}\varepsilon_2 + \sin\mu_{13}\sqrt{\beta_3}\varepsilon_3 = 0,$$
$$\sqrt{\beta_1}\varepsilon_1 + \cos\mu_{12}\sqrt{\beta_2}\varepsilon_2 + \cos\mu_{13}\sqrt{\beta_3}\varepsilon_3 = 0,$$
$$(\cot\pi Q\cos\mu_{1\mathrm{pue}} + \sin\mu_{1\mathrm{pue}})\sqrt{\beta_1}\varepsilon_1 + (\cot\pi Q\cos\mu_{\mathrm{pue}2} + \sin\mu_{\mathrm{pue}2})\sqrt{\beta_2}\varepsilon_2 + [\cot\pi Q\cos(\mu_{\mathrm{pue}2}+\mu_{23}) + \sin(\mu_{\mathrm{pue}2} + \mu_{23})]\sqrt{\beta_3}\varepsilon_3 = \frac{2x_m}{\sqrt{\beta_{\mathrm{pue}}}}, \tag{8.1.2}$$

where

$$\varepsilon = \frac{B_c\Delta s}{B\rho} \tag{8.1.3}$$

is the bump (kick) in the correcting dipole, and

$$\mu_{12} = \int_{s1}^{s2}\frac{ds}{\beta(s)} = Q(\varphi_2-\varphi_1) \tag{8.1.4}$$

is the betatron phase advance.

For the special case of a BPM situated very close to the central corrector, i.e., when $\mu_{\mathrm{pue}2}\approx 0$, the system (8.1.2) reduces to

$$\varepsilon_1 = -\frac{\eta_m}{\sqrt{\beta_1}\sin\mu_{12}},$$
$$\varepsilon_2 = \frac{\sin(\mu_{12}+\mu_{23})}{\sin\mu_{12}\sin\mu_{23}}\frac{\eta_m}{\sqrt{\beta_2}\sin\pi Q},$$
$$\varepsilon_3 = -\frac{\eta_m}{\sqrt{\beta_3}\sin\mu_{23}}. \tag{8.1.5}$$

In the completely symmetrical case, $\mu_{12}=\mu_{23}=\mu$, $\mu_{\mathrm{pue}}=0$, we have

$$\sqrt{\beta_1}\varepsilon_1 = \sqrt{\beta_2}\varepsilon_2 = \frac{x_m}{\sqrt{\beta_{\mathrm{pue}}}\sin\mu},$$
$$\sqrt{\beta_2}\varepsilon_2 = -\frac{2x_m}{\sqrt{\beta_{\mathrm{pue}}}}\cot\mu. \tag{8.1.6}$$

Besides using the expression (3.9) about the closed orbit, we can obtain the system (9.1.2) in another way—by using the beam transport matrix between two arbitrary points s_1 and s_2 in Twiss's form:

$$M_{12} = \begin{bmatrix} \sqrt{\frac{\beta_2}{\beta_1}}(\cos\mu_{12}+\alpha_1\sin\mu_{12}) & \sqrt{\beta_1\beta_2}\sin\mu_{12} \\ \frac{1}{\sqrt{\beta_1\beta_2}}[-\sin\mu_{12}-\alpha_1\alpha_2\sin\mu_{12}+(\alpha_1-\alpha_2)\cos\mu_{12}] & \sqrt{\frac{\beta_1}{\beta_2}}(\cos\mu_{12}-\alpha_2\sin\mu_{12}) \end{bmatrix}. \tag{8.1.7}$$

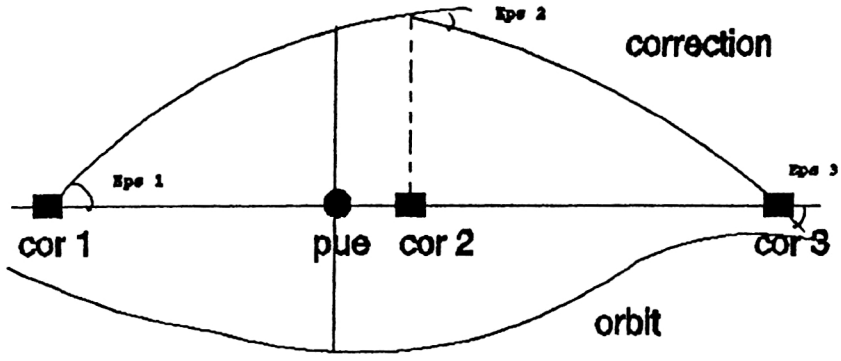

FIG. 3. Beam-bump correction.

Let M^1 be the transport matrix from CD_1 to PUE; M^2, from CD_1 to CD_2; and M^3, from CD_2 to CD_3. Then from (9.1.1) it follows that

$$\varepsilon_1=\frac{x_m}{m_{12}^1},$$

$$\varepsilon_2=-\frac{m_{11}^3 m_{12}^2}{m_{12}^3}\varepsilon_1-m_{22}^2\varepsilon_1,$$

$$\varepsilon_3=-m_{12}^2 m_{21}^3\varepsilon_1-m_{22}^3(m_{22}^3\varepsilon_1^2+\varepsilon_2). \tag{8.1.8}$$

The optimum phase distance between the correctors in the triplet is $\mu\approx\pi/2$ when $\varepsilon_2\approx 0$.

8.2. Generalizations of the beam-bump correction method

As a rule, the magnetic structure of storage rings and colliders is irregular. However, if the phase distance between BPM and CD_2 is large (Fig. 3), the beam-bump method correcting the orbit in the monitor close to zero will increase the orbit deviation in CD_2. This happens when correctors and monitors are situated irregularly (because of lack of space) or when the betatron phase changes quite quickly (in areas with small β). That is why some improvements of the beam-bump method have been put forward. Another reason for such improvements is the fact that in the beam-bump method $\varepsilon_2 \sim 1/\sin\mu$, and in the case when $\mu\approx k\pi$ $(k=1,2,\ldots)$ quite large corrector kicks will be necessary for a full correction to be fulfilled.

In Ref. 15 groups of N monitors and $K=N+2$ correctors are used. The BPMs are considered to be situated close to the corresponding internal correctors.

The kicks of the internal correctors are determined by

$$R\boldsymbol{\varepsilon}_{\text{int}}=-\mathbf{x}_{\text{pue}}, \tag{8.2.1}$$

where the matrix R is given by

$$R_{ij}=\frac{\sqrt{\beta_i\beta_j}\,\sin_{jk}\,\sin\mu_{1i}}{\sin\mu_{1k}}. \tag{8.2.2}$$

In (8.2.2), μ_{1i} is the phase advance between the first corrector and the ith internal corrector; μ_{jk} is the phase advance between the jth internal corrector and the last corrector; μ_{1k} is the phase advance for the whole group of correctors.

Equation (8.2.1) is a generalization of (8.1.2) for the case of $N+2$ correctors and N BPMs.

The two end kicks are determined in such a way as to keep the orbit outside the group unchanged:

$$\sqrt{\beta_1}\varepsilon_1=\frac{x_{\text{pue1}}}{\sqrt{\beta_{\text{pue1}}}\,\sin\mu_{1\text{pue1}}}, \tag{8.2.3}$$

$$\sqrt{\beta_k}\varepsilon_k=\frac{x_{\text{pue}N}}{\sqrt{\beta_{\text{pue}N}}\,\sin\mu_{\text{pue}Nk}}. \tag{8.2.4}$$

In the computer program PETROC (an improved CERN variant of HERA's PETROS) another beam-bump improvement is made.[16,17] One can use K correctors, where $k=0,3,4,5$. Between the correctors $N\leqslant 2k$ beam position monitors are situated. No more than two BPMs can be placed between two correctors. Only three correctors are active (with nonzero kicks)—the first, the second, and the last. Let ε be the kick in the first active corrector.

For the orbit to be unchanged, the other two active kicks have to be εr_2 and εr_k, where

$$r_2=-\sqrt{\frac{\beta_1}{\beta_2}}\frac{\sin\mu_{1k}}{\sin\mu_{2k}}, \tag{8.2.5}$$

$$r_K=\sqrt{\frac{\beta_1}{\beta_k}}\frac{\sin\mu_{12}}{\sin\mu_{2k}}. \tag{8.2.6}$$

The kick ε in the first corrector is determined by the minimum of the function

$$q=\sum_{\text{BPMs}}(x_i+x_{\text{pue}i})^2+w^2\beta_1\beta_2\sin^2\mu_{12}\varepsilon^2. \tag{8.2.7}$$

In (8.2.7), $x_{\text{pue}i}$ is the measured orbit deviation and x_i is the orbit deviation due to the correctors; w is a weight. The second term in (8.2.7) limits the corrector strengths.

9. METHODS WITH ORBIT DEVIATION COMPENSATION

9.1. Iterative beam-bump method

If we move systematically along the accelerator circumference, the local beam-bump method can be used as a method for general correction. Each corrector works once as a first corrector in the corrector triplet (Fig. 3), once as a second, and once as a third corrector.

In this straightforward improvement, a little problem is hidden. In the beam-bump method, the correctors of the triplet are tuned to cancel the orbit in the monitor which is situated as a rule near the middle corrector. Correctors affect the orbit locally; i.e., the impact of the correctors is equal to zero outside the triplet. But as we pointed out above, in the iterative beam-bump the successive triplets overlap slightly (they have two common correctors). Thus, each corrector triplet will cause a small orbit distortion in the next (belonging to the next triplet) beam position monitor. The last corrector triplet consisting of correctors $k-1$, k, and 1 will reduce the orbit deviation in the kth monitor to zero, but will distort the orbit slightly in the first monitor.

To avoid this trouble, one more turn of successive beam-bump corrections along the accelerator will be necessary.

9.2. Least-squares method

The method can be characterized as a single-step correction method.

The orbit is considered only at the points where beam position monitors are situated and is characterized by the state vector $\boldsymbol{\eta}^c$ corresponding to the corrections:

$$\boldsymbol{\eta}=\boldsymbol{\eta}^{(B+g)}+\boldsymbol{\eta}^c=\boldsymbol{\eta}^{(B+g)}+A\,\boldsymbol{\delta}^c. \tag{9.2.1}$$

To write (9.2.1) we have used Eq. (3.9), where $\boldsymbol{\delta}^C$ denotes the generalized corrections (3.11), and the fact that the $\boldsymbol{\eta}^{B+g}$ are known from the BPM readings.

In the least-squares (LSQ) method the following goal function is minimized:[8,18–20]

$$q=\sum_{i=1}^{N}\eta_i^2=\boldsymbol{\eta}^T\boldsymbol{\eta}\rightarrow\min. \tag{9.2.2}$$

It is important to note that as a rule the number of correctors k is less than the number of detectors n, i.e., $K<N$. It can be demonstrated that the minimum of (9.2.2) occurs when the corrector strengths are

$$\boldsymbol{\delta}^c_{\text{opt}}=-(A^TA)^{-1}A^T\boldsymbol{\eta}^{(B+g)}. \tag{9.2.3}$$

The minimum value of q (the so-called residual sum of squares) is

$$q_{\min}=\boldsymbol{\eta}^{(B+g)T}\boldsymbol{\eta}^{(B+g)T}A\,\boldsymbol{\delta}^c_{\text{opt}}. \tag{9.2.4}$$

9.3. Fast realization of the LSQ correction

A fast realization of the LSQ correction method was proposed in Ref. 15. The idea is to optimize the orbit at each iteration, using only one corrector. Having tried in this way all available correctors, we choose for further use the one for which the residual sum of squares of the orbit deviations is the smallest.

At the zeroth iteration we begin with the orbit measured by the BPMs:

$$\boldsymbol{\eta}^{(0)}=\boldsymbol{\eta}^{(B+g)}. \tag{9.3.1}$$

Let $\boldsymbol{\eta}^{(n-1)}$ be the orbit that has been optimized at the $(n-1)$th iteration. At each step at the nth iteration, we use only single correctors. Therefore

$$\boldsymbol{\eta}_j^{(n)}=\boldsymbol{\eta}^{(n-1)}+\delta_j^{(n)}\mathbf{a}_j, \tag{9.3.2}$$

where $\delta_j^{(n)}$ is the generalized strength of the jth corrector at the nth iteration; $\mathbf{a}_j$ is the jth column of the matrix (3.10).

Following the LSQ algorithm, we minimize

$$q_j^{(n)}=\min_{\delta_j^{(n)}}(\boldsymbol{\eta}_j^{nT}\boldsymbol{\eta}_j^{(n)}). \tag{9.3.3}$$

Thus, at the nth iteration, we obtain k values $q_j^{(n)}$ ($j=1,2,\ldots,k$), one for each corrector. From these values $q_j^{(n)}$ we choose the smallest one. The corresponding corrector will be the optimum corrector at the nth iteration. Its optimum strength $\delta_{j(n)}$ will be the correction used at the nth iteration. Therefore our goal function is

$$q_j^{(n)}\rightarrow\min_j. \tag{9.3.4}$$

In Ref. 15 it was proved that this iterative procedure is convergent and that the corrector strengths found at different steps are added.

9.4. The MICADO algorithm

The MICADO algorithm for closed orbit correction was put forward by Autin and Marti in Ref. 21 and is realized in many accelerator design programs, for example in MAD.[22] This algorithm can be seen as an improvement of the LSQ method.

MICADO is an iterative algorithm. In the first iteration only single correctors are used. The goal function is the same as in the LSQ method:

$$q_j^{(1)}=\min_{\delta j}(\boldsymbol{\eta}_j^{(1)T}\boldsymbol{\eta}_j^{(1)}), \tag{9.4.1}$$

where

$$\boldsymbol{\eta}_j^{(1)}=\boldsymbol{\eta}^{(B+g)}+a\,\boldsymbol{\delta}_j^{(1)}, \tag{9.4.2}$$

in which $\boldsymbol{\eta}^{(B+g)}$ is the vector of BPM readings and $\boldsymbol{\delta}_j^{(1)}$ is the vector of the generalized corrections (3.11). As in the first iteration, we use only single correctors:

$$\boldsymbol{\delta}_j^{(1)}=(0,\ldots,\delta_j^{(1)},\ldots,0). \tag{9.4.3}$$

Therefore, in the first iteration we must solve k equations with one unknown value $\delta_j^{(1)}$ ($j=1,2,\ldots,k$).

Next we find the number j^* of the corrector from among all k correctors for which $q_j^{(1)}$ (9.4.1) has the smallest value $q_{j^*}^{(1)}$.

It is the j^*th corrector that will be used in the second iteration together with each one of the other $k-1$ correctors. We minimize again the square function:

$$q_j^{(2)}=\min_{\delta j}(\boldsymbol{\eta}_j^{(2)T}\boldsymbol{\eta}_j^{(2)}). \tag{9.4.4}$$

However, now the correction vector $\boldsymbol{\delta}_j^{(2)}$ is

$$\boldsymbol{\delta}_j^{(2)}=(0,\ldots,\delta_{j^*},\ldots,\delta_j,\ldots,0), \tag{9.4.5}$$

i.e., it has two nonzero components—one (δ_{j^*}) is the strength of the j^*th corrector (remember that this number was found and fixed in the previous step), and the other is the strength of any other corrector. Thus, in the second iteration we must solve $k-1$ systems of two equations with two unknowns δ_{j^*} and δ_j.

Following the first iteration idea, we now seek the number j^{**} of the additional corrector for which $q_j^{(2)}$ has the smallest value. Thus, after the second iteration we have fixed the numbers j^* and j^{**} of two optimum correctors.

We continue this iteration procedure, adding one new corrector at every step, until the sum of the corrected orbit deviations becomes less than a given small value ε. Furthermore, the method finds the smallest number of correctors satisfying this condition.

9.5. The Hereward–Bacconier method

In this correction method the following goal function is introduced:[23–26]

$$q=\gamma\sum_{i=1}^{N}\eta_i^2+(1-\gamma)\sum_{j=1}^{K}\delta_j^{c2}\rightarrow\min. \tag{9.5.1}$$

In (9.5.1), γ is a parameter, with $0<\gamma<1$.

The first sum in (9.5.1) minimizes the orbit deviations in the BPMs, and the second one restricts the strengths of the correctors, limiting in this way the influence of the high harmonics of the correction. We saw that from the readings in N BPMs the amplitudes of the first $N/2$ orbit harmonics can be calculated. These $N/2$ orbit harmonics are compensated by the corresponding harmonics of the correction. The higher harmonics of the correction remain uncompensated. As they distort the orbit additionally, it is useful to restrict their influence.

The optimum value of the parameter γ is determined either experimentally or by computer simulations.

One can write

$$q=\gamma\sum_{i=1}^{N}\left(\sum_{j=1}^{K}a_{ij}\delta_j^c+\eta_i^{(B+g)}\right)^2+(1-\gamma)\sum_{j=1}^{K}\delta_j^{c2}. \tag{9.5.2}$$

The necessary conditions for q to have a minimum are

$$\frac{\partial q}{\partial\delta_j^c}=2\gamma\sum_{i=1}^{N}a_{ij}\sum_{p=1}^{K}(a_{ip}\delta_p^c+\eta_i^{(B+g)})+2(1-\gamma)\delta_p^c$$
$$=0\quad(j=1,2,\ldots,k). \tag{9.5.3}$$

These conditions can be written in matrix form:

$$(\gamma A^TA+(1-\gamma)E)\boldsymbol{\delta}^c=-\gamma A^T\boldsymbol{\eta}^{(B+g)}. \tag{9.5.4}$$

9.6. LSQ method with singular-value decomposition

According to linear algebra,[27] any $N\times K$ matrix A with $N\geq K$ can be represented as a product of three matrices in the form

$$A=UWV^T, \tag{9.6.1}$$

where U is an $N\times K$ unitary matrix ($U^TU=UU^T=E$), W is a $K\times K$ diagonal matrix with positive or zero diagonal elements ($w_i\geq 0$, called singular values), and V is a $K\times K$ unitary matrix ($V^TV=VV^T=E$). This is known as the SVD of the matrix A.

If the matrix A is square (equal numbers of monitors and correctors) and some of the w_i are equal to zero, A will be singular; if some of the w_i are nonzero but very small, A will be ill-conditioned.

If $N>K$ (which is the usual case for accelerators), the system of equations for the corrector strengths

$$A\boldsymbol{\delta}^{BC}=-\boldsymbol{\eta} \tag{9.6.2}$$

will be overdetermined.

Even in these two cases the SVD of A provides a "reasonable" solution of (9.6.2) (in the LSQ sense).

It has been proven[27] that the vector

$$\boldsymbol{\delta}^{BC}=-V[\mathrm{diag}(1/w_i)]U^T\boldsymbol{\eta}, \tag{9.6.3}$$

where for all w_i that are equal to zero or very small the corresponding diagonal elements $(1/w_i)$ in the second matrix of the matrix product (9.6.3) are replaced by zero ($\infty\rightarrow 0!$), satisfies

$$|A\boldsymbol{\delta}^{BC}+\boldsymbol{\eta}|^2\rightarrow\min. \tag{9.6.4}$$

If $U_{(i)}$ are the columns of the matrix U (N vectors) and $V_{(i)}$ are the columns of the matrix V (K vectors), then

$$\boldsymbol{\delta}^{BC}=-\sum_i\left(\frac{U_{(i)}\boldsymbol{\eta}}{w_i}\right)V_{(i)}, \tag{9.6.5}$$

i.e., the corrector strengths are linear combinations of the vectors $V_{(i)}$ with coefficients equal to the dot products of the vectors $U_{(i)}$ with the measured orbit $\boldsymbol{\eta}$ weighted by the singular values w_i.

In order to reduce the corrector strengths, one can replace $1/w_i$ by zero not only for zero (or very small) w_i but also when $w_i<\varepsilon$, where ε is chosen to avoid power-supply saturation.[28]

The LSQ correction method with SVD of the response matrix A is coded in the computer program ORBIT[29] for orbit simulation and correction and has been used in the x-ray ring NSLS and in SPEAR.[30]

9.7. Eigenvector correction

In the LSQ correction method we must invert the matrix A^TA (9.2.3). The matrix A^TA is a square ($K\times K$) real symmetric matrix, i.e., a Hermitian matrix; i.e., there exists a transformation $\Lambda=T^{-1}(A^TA)T$ to a square diagonal matrix Λ. The diagonal elements λ_i of the matrix Λ are the eigenvalues of the matrix A^TA, and the transformation matrix $T=(\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_K)$ is formed by the eigenvalues $\mathbf{X}_i$ of the matrix A^TA.

It follows from this transformation and from (9.2.3) that[8]

$$\delta_i^{BC}=-\sum_{p=1}^{K}\sum_{q=1}^{K}\sum_{r=1}^{K}\frac{X_{p,i}X_{p,q}}{\lambda_p}A_{rq}\eta_r. \tag{9.7.1}$$

Let us represent the corrector vector $\boldsymbol{\delta}$ in the basis $\mathbf{X}_l$ ($l=1,\ldots,K$):

$$\boldsymbol{\delta}=\sum_l C_l\mathbf{X}_l. \tag{9.7.2}$$

Then we have for the residual sum of squares (9.2.4) the expression

$$q_{\min}=\boldsymbol{\eta}^T\boldsymbol{\eta}+\sum_l \lambda_l C_l \boldsymbol{\eta}^T\mathbf{X}_l. \tag{9.7.3}$$

The eigenvectors with small eigenvalues will make no significant contribution to the correction and may be neglected, thus reducing the corrector strengths.

9.8. Dynamical correction algorithm

The algorithms described above strive to compensate orbit deviations at the points where BPMs are situated by means of some number of orbit correctors. As a result of the orbit correction, the corrected orbit will have approximately zero deviation at the BPMs. However, between the monitors the corrected orbit will have nonzero deviations. During the computer simulation we noticed that in some particular error distributions the deviations of the corrected orbit at some points are out of control. In general it is important to correct the orbit over the whole accelerator ring and not only at the BPMs.

Emphasizing this fact, we will choose the criterion for correction quality as a functional:[31]

$$q=\frac{1}{2\pi}\int_0^{2\pi}\eta^2(\varphi)d\varphi. \tag{9.8.1}$$

But the orbit $\eta(\phi)$ is a random function of the generalized azimuth. This requires that we improve the criterion (9.8.1), taking the mean of the functional:

$$q=M\left(\frac{1}{2\pi}\int_0^{2\pi}\eta^2(\varphi)d\varphi\right). \tag{9.8.2}$$

Equation (9.8.2) is the final form of the correction-quality criterion in the DINAM algorithm.

One can write

$$\begin{aligned}\frac{1}{\pi}\int_0^{2\pi}\eta^2(\varphi)\delta\varphi=\frac{u_0}{2}+\sum_{k=1}^{\infty}(u_k^2+v_k^2)=&\left[\frac{u_0^{(B+g)^2}}{2}\right.\\&\left.+\sum_{k=1}^{\infty}(u_k^{(B+g)^2}+v_k^{(B+g)^2})\right]\\&+\left[u_0^{(B+g)}u_0^{Bc}+\sum_{k=1}^{\infty}(2u_k^{(B+g)}u_k^{Bc}\right.\\&\left.+2v_k^{(B+g)}v_k^{Bc})\right]+\left[\frac{u_0^{Bc^2}}{2}+\sum_{k=1}^{\infty}(u_k^{Bc^2}\right.\\&\left.+v_k^{Bc^2})\right]=\sum_1+\sum_2+\sum_3.\end{aligned} \tag{9.8.3}$$

In (9.8.3), u_k, v_k are the Fourier coefficients of the orbit; $u_k^{(B+g)}$ and $v_k^{(B+g)}$ are those caused by the dipole and the quadrupole errors; and u_k^{Bc} and v_k^{Bc} are those caused by the correctors.

Suppose that we have $2N$ beam position monitors, M dipoles, L quadrupoles, and K correcting dipoles. As a result of the orbit measurement by the BPMs, we will know the orbit distortions at $2N$ points. However, the total number of perturbations $(M+L)$ is, as a rule, much larger than the number of BPMs. Therefore, we cannot solve (3.9) for the perturbations. From this system of equations only $2N$ perturbations can be expressed in terms of the readings in the BPMs and other perturbations. In other words, we have a case of correction with uncertainty.

Let η_i $(i=1,2,\ldots,2N)$ be the orbit distortions at the BPMs. Then we can write

$$\eta_i=\frac{u_0}{2}+\sum_{k=1}^{\infty}(u_k\cos k\varphi_i+v_k\sin k\varphi_i)\quad(i=1,2,\ldots,2N). \tag{9.8.4}$$

From these $2N$ equations, the first N Fourier coefficients $u_0,u_1,\ldots,u_N,\ldots,v_{(n-1)}$ can be expressed in terms of the orbit distortions η_i and higher orbit harmonics. Assuming that the BPMs are placed uniformly along the accelerator ring, the following relations can be deduced from (9.8.4):

$$\begin{aligned}u_0&=U_0-2\sum_{j=1}^{\infty}u_{2Nj}=U_0-S_0,\\u_k&=U_k-\sum_{j=1}^{\infty}(u_{2Nj-k}+u_{2Nj+k})=U_k-S_k,\\v_k&=V_k-\sum_{j=1}^{\infty}(-v_{2Nj-k}+v_{2Nj+k})=V_k-R_k,\end{aligned} \tag{9.8.5}$$

where

$$\begin{aligned}U_0&=\frac{1}{N}\sum_{i=1}^{2N}\eta_i,\\U_k&=\frac{1}{N}\sum_{i=1}^{2N}\eta_i\cos k\varphi_i\quad(k=1,2,\ldots,N),\\V_k&=\frac{1}{N}\sum_{i=1}^{2N}\eta_i\sin k\varphi_i\quad(k=1,2,\ldots,n-1)\end{aligned} \tag{9.8.6}$$

are Bessel coefficients.

Making use of the relations (9.8.5), one can write, for the sums in (9.8.3),

$$\begin{aligned}\sum_1=&\left[\frac{U_0^2}{2}+\sum_{k=1}^{n-1}(U_k^2+V_k^2)+\frac{U_N^2}{2}\right]+\left[-U_0S_0\right.\\&\left.+\sum_{k=1}^{N-1}(-2U_kS_k-2V_kR_k)-U_NS_N\right]+\left[\frac{S_0^2}{2}\right.\\&\left.+\sum_{k=1}^{N-1}(S_k^2+R_k^2)+S_N^2-\frac{U_N^2}{4}+V_N^{(B+g)^2}\right.\\&\left.+\sum_{k=N+1}^{\infty}(u_k^{(B+g)^2}+v_k^{(B+g)^2})\right]=\sum_a+\sum_b+\sum_c,\end{aligned} \tag{9.8.7}$$

$$\begin{aligned}\sum_2=&\left[U_0u_0^{Bc}+2\sum_{k=1}^{n-1}(U_ku_k^{Bc}+V_kv_k^{Bc})+U_Nu_N^{Bc}\right]\\&+\left[2v_N^{B+g}v_N^{Bc}-u_0^{Bc}S_0-2\sum_{k=1}^{n-1}(u_k^{Bc}S_k+v_k^{Bc}R_k)\right.\end{aligned}$$

$$-2u_N^{Bc}S_N\Big]+\Big[2\sum_{k=N+1}^{\infty}(u_k^{(B+g)}u_k^{Bc}+v_k^{(B+g)}v_k^{Bc})\Big]$$

$$=\sum_d+\sum_e+\sum_f. \tag{9.8.8}$$

Using the fact that the higher orbit harmonics are statistically independent and have zero mean, we obtain

$$M\Big(\sum_d\Big)=M\Big(\sum_e\Big)=M\Big(\sum_f\Big)=0. \tag{9.8.9}$$

Using the Fourier expansion (3.13) and the relations (3.14) between the orbit and the perturbation harmonics, one can also obtain

$$M\Big(\sum_c\Big)=4\sum_{k=N+1}^{\infty}\left(\frac{Q^2}{Q^2-k^2}\right)^2 D, \tag{9.8.10}$$

where D is the variance of the perturbation harmonics:

$$D=D(f_k)=D(g_k). \tag{9.8.11}$$

Because the correcting dipoles are short in comparison with the accelerator circumference, one can reduce the integrals in the expressions for the corrector Fourier harmonics to sums and write

$$\sum_d=\sum_{p=1}^{2N}\sum_{q=1}^{K}A_{pq}\eta_p\delta_q^{Bc}, \tag{9.8.12}$$

where

$$A_{pq}=\frac{1}{N}\Big(c_0+2\sum_{k=1}^{n-1}c_k\cos k(\varphi_p-\varphi_q)+c_N\cos N\varphi_p\cos N\varphi_q\Big),$$

$$c_k=\frac{2Q\sin\pi Q}{\pi^2(Q^2-k^2)}. \tag{9.8.13}$$

In the same way, one can prove that

$$\sum_3=\sum_{q=1}^{k}\sum_{r=1}^{k}B_{qr}\delta_q^{Bc}\delta_r^{Bc}, \tag{9.8.14}$$

where

$$B_{qr}=\frac{c_0^2}{2}+\sum_{k=1}^{\infty}c_k^2\cos K(\varphi_q-\varphi_r)=\cos Q|\varphi_q-\varphi_r|+\frac{\sin\pi Q}{\pi Q}\cos(\pi-|\varphi_q-\varphi_r|)Q-\frac{|\varphi_q-\varphi_r|\sin\pi Q}{\pi}\sin(\pi-|\varphi_q-\varphi_r|)Q. \tag{9.8.15}$$

Finally, in the applied harmonic analysis it is proved that

$$\sum_a=\frac{1}{N}\sum_{i=1}^{2N}\eta_i^2. \tag{9.8.16}$$

Summarizing all the above results for the quality criterion (9.8.2), we get

$$2q=\frac{1}{N}\sum_{i=1}^{2N}\eta_i^2+\sum_{p=1}^{2N}\sum_{q=1}^{k}A_{pq}\eta_p\delta_q^{Bc}+\sum_{q=1}^{k}\sum_{r=1}^{k}B_{qr}\delta_q^{Bc}\delta_r^{Bc}+4\sum_{k=n+1}^{\infty}\left(\frac{Q^2}{Q^2-k^2}\right)D. \tag{9.8.17}$$

The coefficients A_{pq} and B_{qr} are given by Eqs. (9.8.13) and (9.8.15), respectively. In these formulas, φ_p are the azimuths of the BPMs, and φ_q and φ_r are the azimuths of the correcting dipoles.

The strengths of the correcting dipoles are determined by the condition for a minimum of q to occur:

$$2\sum_{p=1}^{k}B_{sp}\delta_p^{Bc}=-\sum_{p=1}^{2N}A_{ps}\eta_p\quad(s=1,2,\ldots,k). \tag{9.8.18}$$

Introducing matrices $A=\{A_{ij}\}$ and $B=\{B_{ij}\}$ and the vectors of the corrections $\boldsymbol{\delta}^{Bc}$ and BPM readings $\boldsymbol{\eta}$, Eq. (9.8.18) can be written in the following matrix form:

$$2B\boldsymbol{\delta}^{Bc}=-A^T\boldsymbol{\eta}. \tag{9.8.19}$$

Let us introduce the matrix

$$R=-\frac{1}{2}B^{-1}A^T. \tag{9.8.20}$$

The matrix R depends only on the azimuths of the BPMs and on the correctors, and for the given accelerator it can be calculated prior to the correction. Then the required strengths of the correcting dipoles will be determined by the matrix expression

$$\boldsymbol{\delta}^{Bc}=R\boldsymbol{\eta}. \tag{9.8.21}$$

Thus, the algorithm is relatively fast. The computer simulations showed that it works reliably and is free from the undesirable effects mentioned at the beginning of this section.

10. ORBIT CORRECTION WITH NEURAL NETWORKS

Artificial neural networks (ANNs)[32] with ability to tune themselves according to the output errors are well suited for on-line orbit correction. The relationship between corrector kicks and orbit displacements in BPMs is in general nonlinear, owing to the presence of nonlinear elements in the machine. The orbit fluctuates in time. Both of these circumstances are well treated by ANNs, which have features for solving nonlinear problems and self-teaching.

As it is not possible to go into detail in this paper, we will restrict ourselves to the main principles of ANN correction.

An ANN consists of neurons arranged in layers, with directed and weighted connections between them (Fig. 4).

In the forward propagation of the signals (from input to output) each neuron processes its input signals s_j and produces an output signal s_i according to the relation

$$s_i=f\Big(\sum_{j=i}T_{ij}s_j\Big), \tag{10.1}$$

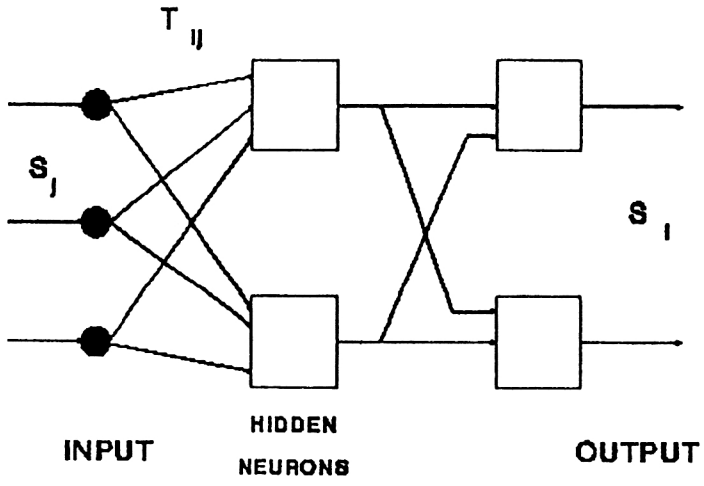

FIG. 4. Artificial neural network.

where the T_{ij} are synapse weights.

The function $f(x)$ in (10.1) (the so-called action function) is usually a step function; thus, if the weighted sum of the input signals exceeds the threshold for the neuron, the neuron "fires" and produces an output signal.

In the backward propagation of the signals, first of all the output errors ε_i of the ANN are calculated:

$$\varepsilon_i = z_i - s_i, \tag{10.2}$$

where the z_i are the "ideal" output signals.

Let B_{ij} be the blames—the degree of responsibility that each input signal has for the output error:

$$B_{ij} = \frac{T_{ij} s_j}{\Sigma_{j=i} T_{ij} s_j}. \tag{10.3}$$

Finally, the weights are corrected according to the relation

$$T_{ij}^{\text{new}} = T_{ij}^{\text{old}} + k B_{ij} \varepsilon_i, \tag{10.4}$$

where k is a coefficient.

Before use, an ANN must be trained. To do this we apply a set of input signals for which the "ideal" outputs are known, record the output errors, and tune the weights according to (10.4). After many training cycles the desired accuracy of the output can be reached.

The use of ANNs for orbit correction is quite straightforward. The input signals are the measured orbit displacements $\mathbf{X}$ in the BPMs, and the output signals are the corrector kicks $\boldsymbol{\varepsilon}$. There is one input and one output neuron for each BPM and corrector.

In the training stage the applied input signals $\mathbf{X}$ are determined by

$$\mathbf{X} = A\boldsymbol{\varepsilon}, \tag{10.5}$$

where A is the response matrix, whose elements can be calculated by the machine model or can be measured, which is more accurate, and $\boldsymbol{\varepsilon}$ are random kicks.

After training and attainment of the desired accuracy of the output, the ANN can be used for on-line orbit correction.

During the machine operation a continuous process of retraining (fine tuning) goes on. The detected orbit errors are fed back, and the weights are recalculated (adaptive correction).

ANNs have been used for orbit correction in the NSLS VUV and x-ray storage rings.[33,34]

Neural networks have been simulated by means of SNNS computer simulator.[35] A three-layer shortcut connected network (all neurons are connected to each other) and Quickprop training strategy chose the best results. After 2200 training cycles the ANN was able to correct the orbit to 44 μm maximum deviation.

11. EXPERT SYSTEMS

Knowledge-based expert systems differ from conventional computer programs in their intensive use of intuitive and empirical rules which, together with facts about the task, form a so-called knowledge base.[36] The control strategy (the order in which rules are applied) is determined by an inference engine (Fig. 5).

The expert-system approach to closed orbit correction is being developed at CERN by Brandt and Verdier.[37,38]

The expert system is based on Guignard's fitting method (Sec. 7.2) of searching for field and alignment errors. It can predict and estimate the location of large field defects and check the correctness of the BPM reading. A convergence test is used as a sensitive means of detecting whether all the errors have been identified.

The main program is written in Prolog, while the numerical subroutines are in Pascal.

The expert system has been successfully tested on EAP and LEAR at CERN.

12. ORBIT-CORRECTION FEEDBACK SYSTEMS

In synchrotron-radiation sources (SRS) we need not only closed orbits with small distortions but also highly stable orbits. Orbit stability is a crucial point in achieving low-emittance electron beams and therefore high brightness of

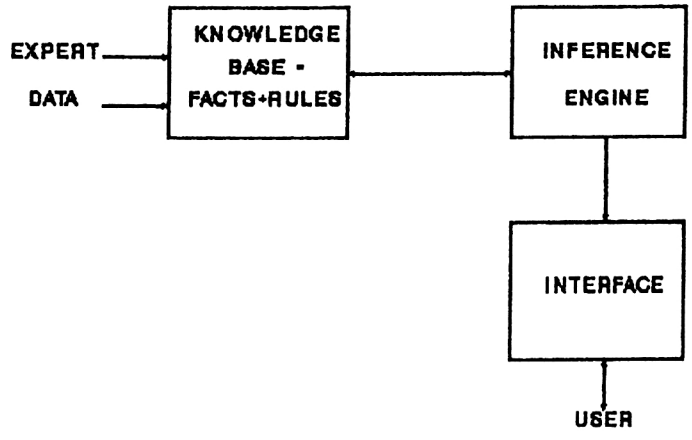

FIG. 5. Block diagram of a knowledge-based expert system.

the photon beams. Orbit correction must be applied dynamically to eliminate fluctuations produced by ground vibrations and magnet power ripples.

In SRS, orbit stability is improved by means of a correction feedback system (Fig. 6).

In general, feedback systems are classified into local and global systems.

In local feedback systems the orbit is locally corrected at the center of an insertion device by three or four magnet bumps.[39]

Global feedback systems can be analog[40] or digital.[41]

In Ref. 40 a feedback system based on Fourier analysis of the orbit is described. As was shown in Sec. 7.1, the orbit Fourier coefficients can be expressed in terms of the matrix equality (7.1.5). This means that a simple linear electronic network can be built to perform on-line Fourier analysis. The input voltages are proportional to the orbit displacements measured by BPMs, and in real time the output voltages are proportional to the Fourier harmonics. Then the corrector strengths are adjusted to cancel these harmonics.

Digital feedback systems[41] make it possible to avoid drift, offset, and temperature problems typical of analog circuits. Orbit data are transferred in digital form by BPM processors distributed around the ring. The digital signal processors then calculate the corrector strengths and control the corrector power supplies.

Digital feedback systems are programmable and therefore more flexible in use.

In principle any orbit-correction algorithm can be used in feedback systems.

13. OPTIMUM POSITIONING OF DIPOLE MAGNETS

A different approach to the closed orbit correction is the optimum positioning of the dipole magnets around the ring. In this approach the dipole magnets are situated around the ring not in an arbitrary way, but following a special strategy.

In the stage of accelerator assembly the field errors in the dipoles are carefully measured and only after that are installed according to a positioning algorithm.

This allows the orbit distortion to be minimized.[42,43]

Suppose that the dipoles are installed around the ring according to the permutation

$$X=(\Delta B_{k_1},\Delta B_{k_2},\ldots,\Delta B_{k_M}),$$

$$k_i\in\{1,2,\ldots,M\},\quad k_i\neq k_j,\quad i\neq j. \tag{13.1}$$

If we choose as a quality criterion the functional

$$q=\frac{1}{2\pi}\int_0^{2\pi}\eta^2(\phi)d\phi\rightarrow\min, \tag{13.2}$$

it can be shown[42] that the following sum must be minimized:

$$\sum_{q=1}^{M}\sum_{r=1}^{M}B_{qr}\Delta B_{k_q}\Delta B_{k_r}\rightarrow\min, \tag{13.3}$$

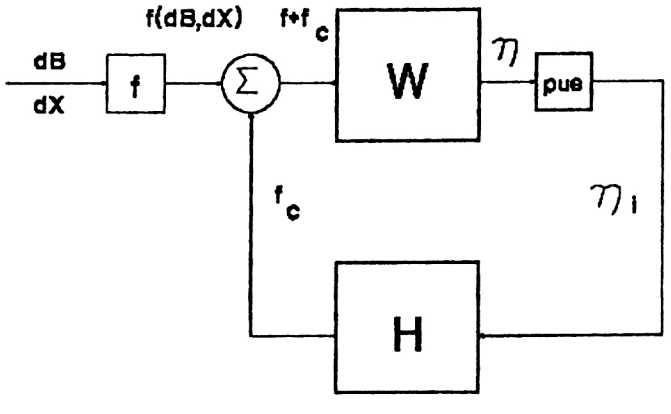

FIG. 6. Orbit-correction feedback system. Here $f(dB,dX)$ is the error function (3.6), W is the operator $1/Q^2(d^2/d\phi^2+Q^2)^{-1}$, and H is the correction operator.

where the coefficients B_{qr} are calculated by the machine model.

Let us introduce the integer variables

$$X_{jk}=\begin{cases}1, & \text{if the } K\text{th dipole is situated}\\ & \text{at the } j\text{th position,}\\ 0, & \text{otherwise,}\end{cases} \tag{13.4}$$

$$\sum_{k=1}^{M} X_{jk}=1, \quad \text{i.e., one dipole at a place,}$$

$$\sum_{j=1}^{M} X_{jk}=1, \quad \text{i.e., each dipole only at one place.} \tag{13.5}$$

From (13.3) the so-called "quadratic assignment problem" of discrete programming follows:

Let us have M dipoles, $k=1,2,\ldots,M$, and M places $j=1,2,\ldots,M$ around the ring. We seek positioning of every dipole at only one place for which

$$\sum_{q=1}^{M}\sum_{r=1}^{M}\sum_{m=1}^{M}\sum_{n=1}^{M}[B_{qr}\Delta B_m \Delta B_n]X_{qm}X_{rn}\rightarrow \min, \tag{13.6}$$

where the X_{jk} are M^2 integer variables (13.4) with the constraints (13.5).

Unfortunately, in cases of large dimension of the task the quadratic assignment problem proves to be a difficult one.[42]

Therefore in Refs. 42 and 43 new approaches to the problem making full use of its specific character have been developed.

The set of dipole errors (13.1) creates a combinatorial space of permutations P_B. The points in this space are all possible permutations X (13.1), and its power is $M!$.

Let us introduce a metric in the P_B space in the following way: the distance $r(X,Y)$ between the points X and Y is assumed to be equal to the minimum number of transpositions (elementary or pair shifts) necessary to bring the point X to the point Y.

Here we will describe only one of the algorithms proposed in Refs. 42 and 43, namely, the algorithm of controlled random search.

The algorithm uses the goal function

$$Q=\max_i|x_i|, \tag{13.7}$$

where x_i is the orbit displacement in the ith BPM.

The logical structure of the algorithm of controlled random search can be described in the following steps.

Controlled random search

Step 1. Choose an arbitrary initial arrangement of the dipoles X_0, i.e., an arbitrary initial point in the combinatorial space P_B. Draw a sphere S_0 centered at the initial point X_0 and having radius equal to R_0 ($R_0<M-1$, and its value is chosen by physical considerations). Choose the convergence parameter ε.

Step 2. Set $i=1$.

Step 3. Choose a random point $X_i\in S_{(i-1)}$, using a uniform probability distribution. Calculate $Q_i=Q(X_i)$.

Step 4. Check whether $Q_i<\varepsilon$. If so, stop the iterations and exit the algorithm. If not, go to step 5.

Step 5. Draw a sphere S_i centered at X_i and having radius R_i:

$$R_i=\delta R_0, \tag{13.8}$$

where

$$\delta=\frac{Q_i}{Q_0}. \tag{13.9}$$

Step 6. Set $i=i+1$ and go back to step 3.

The computational experiments show that the CPU time necessary for optimizing a machine with M dipoles is proportional to $M!$.

Other algorithms for optimum positioning of dipoles can be found in Refs. 42 and 43.

14. FIRST-TURN CORRECTION

If the particles in a circular accelerator are to undergo hundreds of thousands of turns before reaching the final energy of the machine, they must first of all pass the very first turn.

During the assembly and initial tuning of the accelerator much larger errors than the random field and alignment errors mentioned above may occur. Sometimes they are caused by unpredictable mistakes, and there have been several such cases in accelerator practice.

In the presence of large linear errors the center-of-charge trajectory no longer follows the orbit and can have very large deviations. Even worse, the beam can hit the vacuum chamber somewhere and be unable to make a complete turn around the machine.

Launch errors in the injection system may also cause beam loss or, provided they are not so large, harmful coherent oscillations of the beam.

The situation is complicated by displacement errors in the BPMs and by the low resolution of the monitors for a single-pass beam.

Thus, we face the task of threading the beam through the entire turn around the accelerator ring.

Two different approaches are possible.[44]

First, we can try to thread the beam around the ring by using the existing orbit correctors. Correcting algorithms have been developed for first-turn treatment, and some of them will be described in this paper.

In a different approach, we can seek the sources of large errors—dipole magnets with large field errors or highly displaced quadrupoles. After finding candidates for error sources we must carefully examine the corresponding elements. This approach has been successfully applied for beam-line steering and can also be used for the first-turn treatment.

The first-turn correction is closely related to beam-line steering. In fact, before closing the first turn onto the second, the magnetic structure of a circular accelerator can be viewed as a beam line.

Beam losses are especially dangerous in superconducting machines, causing the loss of superconductivity and thus making the tuning process much longer.

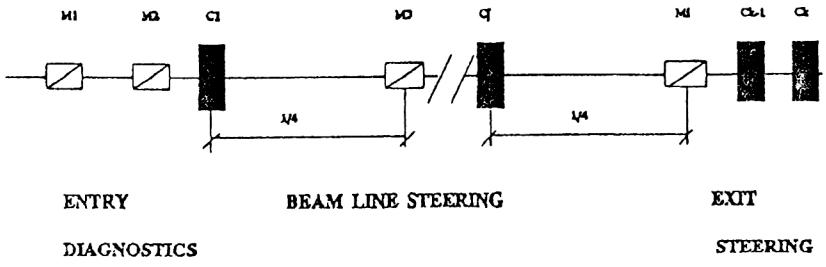

FIG. 7. Beam-line steering system.

In large synchrotrons the distorted orbit amplitude can reach values comparable with the vacuum-chamber radius, mainly owing to misalignments of the quadrupoles. This makes first-turn threading a very important task.

Having corrected the center-of-charge trajectory so that the beam goes through an entire turn around the ring, we must close this trajectory, i.e., make the second turn (and all subsequent turns) coincide with the first turn.

This section presents some algorithms for first-turn steering.

14.1. Beam-threading algorithm

The idea of the method is straightforward.[45,46] The steering system consists of small correcting dipoles and beam position monitors (BPM) situated near them (Fig. 7).

As BPMs, electrostatic pickup electrodes are used.

Because the length of the correcting dipoles is small compared with the accelerator circumference, they can be considered to produce local orbit bumps.

Let

$$\varepsilon_n=\frac{B_{cn}l}{B\rho} \tag{14.1.1}$$

be the kick in the nth corrector, where $B\rho$ denotes the beam rigidity.

If M^{ji} is the transfer matrix between the jth corrector and the ith BPM, we can write

$$x_i=m_{12}^{ji}\varepsilon_j=\sqrt{\beta_j\beta_i}\,\sin\mu_{ji}\varepsilon_j, \tag{14.1.2}$$

where x_i is the deviation in the ith BPM, and we have expressed the matrix elements in terms of the Twiss parameters—the amplitude function $\beta(s)$ and the betatron phase advance $\mu(s)$:

$$\mu(s)=\int_0^s\frac{ds}{\beta(s)}. \tag{14.1.3}$$

Applying the superposition principle, we can write

$$\begin{bmatrix}x_3\\x_4\\\cdots\\x_n\end{bmatrix}=\begin{bmatrix}\sqrt{\beta_1\beta_3}\sin\mu_{13} & 0 & .. & 0\\ \sqrt{\beta_1\beta_4}\sin\mu_{14} & \sqrt{\beta_2\beta_4}\sin\mu_{24} & .. & 0\\ .. & .. & .. & ..\\ \sqrt{\beta_1\beta_n}\sin\mu_{1n} & \sqrt{\beta_2\beta_n}\sin\mu_{2n} & .. & \sqrt{\beta_{k-2}\beta_n}\sin\mu_{k-2n}\end{bmatrix}\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\cdots\\\varepsilon_{k-2}\end{bmatrix}. \tag{14.1.4}$$

The system of equations (14.1.4) can be solved with a trivial forward substitution.

It follows from (14.1.2) that the optimum phase distance between the corrector and the monitor in a corrector–monitor pair is $\pi/2$. In practice it is difficult to fulfill this requirement.

BPMs should be placed near the points where β has a maximum and where we can expect large center-of-charge deviations. This improves the BPM sensitivity.

In order to minimize the corrector strengths, the correcting dipoles should also be placed near the points where β has a maximum.

Whereas in the simplest FODO structure it is easy to follow these rules for situating the horizontal BPMs and correctors near the F quadrupoles and the vertical BPMs and correctors near the D quadrupoles, in more sophisticated magnetic structures with many elements this is almost impossible.

The beam-line steering systems have, as a rule, two additional parts (Fig. 7).

In the entrance of the line two pickups are used to measure the position and the slope of the injected beam. The reliability of the entry-angle reconstruction is greatly improved, however, if the two pickups are placed in the drift space.[45]

At the exit of the line two correctors are used to match the center-of-charge position and slope to those in the following accelerator or target.

14.2. Least-squares algorithm

The least-squares (LSQ) approach has been successfully applied not only as a method for general orbit correction but also for first-turn treatment.[47]

Suppose that we have n BPMs and m correcting dipoles, with $n \geq m$.

If c_k $(k=1,2,\ldots,m)$ are changes in the correcting currents, the theoretical changes in the center-of-charge trajectory x_i^c that they will cause are in general given by

$$x_i^c = \sum_{k=1}^{m} \left(\frac{\partial x_i^c}{\partial c_k} \right) c_k = \sum_{k=1}^{m} F_{ik} c_k . \tag{14.2.1}$$

In practice, owing to the error influence, the measured center-of-charge position x_i^m will not coincide with the design position x_i^d. Let $\Delta x_i = x_i^d - x_i^m$ denote the discrepancy between them.

The task of threading the beam around the entire circumference of the ring can be viewed as a least-squares problem:

$$S^2 = \sum_{i=1}^{n} (\Delta x_i - x_i^c)^2 \rightarrow \min. \tag{14.2.2}$$

The corrector strengths which minimize the discrepancy between the measured and desired trajectories follow from the well-known solution of the LSQ problem:

$$\mathbf{c} = (F^T F)^{-1} F^T \Delta \mathbf{x}. \tag{14.2.3}$$

In the case of equal numbers of correctors and BPMs, when F is a square matrix, (2.2.3) becomes

$$\mathbf{c} = F^{-1} \Delta \mathbf{x}. \tag{14.2.4}$$

For first-turn steering F is a triangular matrix [like (2.1.4)], as a BPM "sees" only those correctors which are placed downstream to it.

The LSQ method has been used in the Tevatron first-turn correction.[47] In superconducting synchrotrons, orbit steering is of special importance because beam losses cause the loss of superconductivity. This means that the accelerator will have about one hour of stoppage time at every tuning step, and this is a serious problem. Automatization of the tuning by means of the LSQ algorithm has made it possible for an orbit close to the design orbit to be established after several iterations of the correction.

Another LSQ algorithm for first-turn steering is described in Ref. 48.

The goal function is defined as

$$\Psi(\varepsilon_j^*) = \sum_{i=1}^{n} \left[(\eta_i^{\text{pue}} - \eta_i^{\text{des}}) + \sum_{j=1}^{k} a_{ij} \varepsilon_j^* \right]^2 \rightarrow \min, \tag{14.2.5}$$

where

$$\varepsilon_j^* = \sqrt{\beta_j^{\text{cor}}}\, \varepsilon_j \tag{14.2.6}$$

are "generalized" kicks.

In (14.2.5), η^{des} denotes the desired center-of-charge trajectory, and n and k are the current numbers of active pickups and correctors. When we lose the beam before the full turn, the current number of active pickups n is less than the total number of pickups N, and the current number of switched-on correctors k is less than the total number of available correctors K.

In practice there are limitations on the maximum allowed kicks in the correcting magnets:

$$|\varepsilon_j| < \Delta. \tag{14.2.7}$$

Thus, we face a constrained optimization problem.[49]

In order to solve this problem, the penalty-function method has been used.

The general idea of the method of penalty functions is to reduce the constrained optimization problem to a series of unconstrained problems. To do this, we add to our goal function (14.2.5) a so-called penalty function $\alpha(\varepsilon_j^*)$, which is chosen in such a way that it will "punish" the function $\Psi(\varepsilon_j^*)$ if the constraints (14.2.7) are violated.

In Ref. 48 the following penalty function was used:

$$\alpha(\varepsilon_j^*) = \sum_{j=1}^{K} \max(0, \varepsilon_j^{*2} - \Delta^2). \tag{14.2.8}$$

The following series of unconstrained optimization problems is solved:

$$\Psi'(\varepsilon_j^*, \mu_k) = \Psi(\varepsilon_j^*) + \mu_k \alpha(\varepsilon_j^*) \rightarrow \min, \tag{14.2.9}$$

where $\mu_k > 0$ $(k=1,2,3,\ldots)$ are parameters, and

$$\lim_{k \rightarrow \infty} \mu_k = \infty. \tag{14.2.10}$$

14.3. Closed bump algorithm

The classical method for global orbit correction by means of local orbit bumps can also be applied for first-turn correction.[49]

The local orbit bump is produced by three correcting dipoles (Fig. 3).

Suppose that n BPMs are placed between the correctors.

Let a be the center-of-charge deviation in the middle corrector. The height a can be used to parametrize the local orbit bump.

A kick ε_i centered at a point i produces changes in the center-of-charge position and angle at another point s given by

$$\Delta x(s) = \sqrt{\beta_i \beta_s}\, \sin \mu_{is} \varepsilon_i ,$$

$$\Delta x'(s) = \sqrt{\frac{\beta_i}{\beta_s}}\, (\cos \mu_{is} - \alpha_s \sin \mu_{is}) \varepsilon_i . \tag{14.3.1}$$

We want to produce a local bump. This means that if the point s is situated anywhere outside the bump, the effects of the three correctors should compensate each other:

$$\sum_{i=1}^{3} \Delta x_i(s) = 0,$$

$$\sum_{i=1}^{3} \Delta x_i'(s) = 0. \tag{14.3.2}$$

It follows from (14.3.1) and (14.3.2) that for s outside the bump

$$\sum_{i=1}^{3} \sqrt{\beta_i} \varepsilon_i \sin \mu_{is} = 0,$$

$$\sum_{i=1}^{3} \sqrt{\beta_i} \varepsilon_i \cos \mu_{is} = 0. \tag{14.3.3}$$

The kick in the first correcting dipole should produce a deviation equal to a in the middle corrector. It should therefore be

$$\varepsilon_1=\frac{a}{\sqrt{\beta_1\beta_2}\sin\mu_{12}}. \tag{14.3.4}$$

Given (14.3.4), Eq. (14.3.3) becomes a system of two equations in two unknowns ε_2 and ε_3.

The solution is

$$\varepsilon_2=\frac{a\sin\mu_{12}}{\beta_2\sin\mu_{12}\sin\mu_{23}}, \tag{14.3.5}$$

$$\varepsilon_3=\frac{a}{\sqrt{\beta_2\beta_3}\sin\mu_{23}}. \tag{14.3.6}$$

In Ref. 50 the following goal function was introduced:

$$G(a)=\sum_{i=1}^{n} w_i^{\text{pue}}(x_i^m-x_i^d+x_i^c)^2+\sum_{j=1}^{3} w_j^{\text{cor}}P_j(\varepsilon_j). \tag{14.3.7}$$

In (14.3.7), x_i^m again denotes the measured deviation in the ith BPM, and x_i^d is the desired deviation; x_i^c is the deviation in the ith BPM due to the three correctors, which can be easily calculated using the transfer matrices between the correctors and the monitor; w_i^{pue} is a weight associated with each monitor, and w_j^c is a weight associated with each corrector; $P(\varepsilon)$ is a penalty function which punishes the goal function if the kicks are too large.

In particular, the penalty function can be taken as

$$P(a)=\begin{cases} k|\varepsilon|, & \text{if } |\varepsilon|>\varepsilon_{\max}, \\ 0, & \text{otherwise,} \end{cases} \tag{14.3.8}$$

where $k>0$ is a large constant.

$G(a)$ is a function of only one independent variable a. It can be easily minimized, for example, by the simplex method.

For the entire circumference of the accelerator ring to be corrected, the above algorithm must be applied iteratively. The ring is divided into overlapping closed bumps, and the correction is repeatedly applied, taking into account the results of the previous steps.

The algorithm is proposed for the SSC and has been largely used for SSC first-turn and global closed orbit simulations.

14.4. Error-finding methods

In the periods of machine assembly, initial tuning, or upgrading, relatively large errors can occur—launch errors in the position and slope of the center of charge of the injected beam, kick errors in the dipoles and misaligned quadrupoles, and focus errors in the quadrupole gradients.

As the strengths of the correcting elements are limited, it may be impossible to correct the center-of-charge trajectory to the required extent. Therefore a different approach to beam steering has been developed.[50,51]

In this approach we seek the sources of large errors instead of trying to correct them. After the positions and magnitudes of such errors have been found one should carefully check the corresponding elements, trying to reveal the physical basis of the errors.

The error-finding approach has been successfully used at SLAC for beam steering in the linear electron–positron collider SLC.[50,51]

Two kinds of computer programs are used to determine the error candidates—modeling programs and error-simulating programs.

The modeling programs give the operator a mathematical model of the accelerator. They receive as input a full description of the accelerator elements (their location and strengths) and produce as output a file listing elements and corresponding transfer matrices and Twiss parameters.

Error-simulating programs calculate the effects on the beam parameters due to specific errors in the elements and generate simulated trajectories. To find the predicted trajectories these programs use the transfer matrices calculated by the modeling programs.

As a modeling program, any lattice design program can be used.

For simplification the error-simulating programs describe errors introducing thin-lens elements in the lattice. To simulate focusing errors, thin-lens quadrupoles are inserted in the lattice quadrupoles, and to simulate kick errors, thin-lens dipoles are inserted either in bending dipoles to describe field errors or in quadrupoles to describe alignment errors.

Two different kinds of error treatment exist:

A. Global search. The method of global error search makes use of a powerful nonlinear optimization package. All the elements are suspected as possible sources of errors. Thin-lens elements describing errors are inserted in the elements. Then the optimization programs are used to find the settings of these thin-lens elements which yield BPM readings matching the measured trajectory. Thin-lens elements which have nonzero strengths are the sites of possible errors.

B. Local search. Only a few of the elements are used as possible error sources. The operator has to make some guesses about the location of the errors and about the error-free regions. A highly developed graphical interface can be very useful in this trial-and-error method. The operator display should be able to show a plot of the measured trajectory, the desired trajectory, and the difference between them. It is clear that an error originates in the region where the differences are large. After determining the error positions the operator tries to adjust the strengths of the corresponding thin-lens elements so that the calculated trajectory lies close to the measured trajectory. Nonlinear optimization programs are used again, but because of the small number of independent variables the optimization problem is much easier to solve.

Several expert systems have been developed to automate the use of beam-line correcting programs and to minimize the time necessary for line commissioning.[52–54]

These are hybrid programs combining traditional expert systems with their capabilities for qualitative reasoning, modeling programs giving a mathematical model of the machine, and optimization programs. Combining numerical algorithms with symbolic reasoning, these hybrid expert sys-

tems systematically perform the specific procedures that a human expert follows in order to correct the beam line.

First of all, they look for error-free regions where the discrepancies between the predicted and the measured trajectories are small. It is assumed that every subregion between two adjacent error-free regions is a possible location of errors and that there is only one error element within each subregion. Then the error-finding procedures described above are used to reveal the existence of beam-line errors.

Frames are usually used for representation of the domain knowledge.

LISP is the preferred language for the expert-system implementation.

*Work supported by the Bulgarian Scientific Foundation, contract F-309.

[1] A. A. Kolomensky and A. N. Levedev, *Theory of Cyclic Accelerators* (North-Holland, Amsterdam, 1966).
[2] H. Bruck, *Accelerateurs Circulaires de Particules* (Press Universitaires de France, Paris, 1966).
[3] P. Strollin, Preprint ISR-TH/68-4, CERN, 1968.
[4] R. Gluckstern, Part. Accel. **8**, 203 (1978).
[5] R. W. Hamming, *Introduction to Applied Numerical Analysis* (McGraw-Hill, New York, 1971).
[6] L. Resegotti, Preprint ISR-MAG/68-30, CERN, 1968.
[7] B. Autin and R. J. Bryant, Preprint ISR-MA/71-36, CERN, 1971.
[8] A. Ando and E. Endo, Preprint 75-4, KEK, 1975.
[9] G. Guignard, Preprint ISR-BOM/80-21, CERN, 1980.
[10] J. L. Waren and P. J. Channell, IEEE Trans. Nucl. Sci. **NS-30**, No. 4 (1983).
[11] P. J. Averill, IEEE Trans. Nucl. Sci. **NS-12**, 899 (1965).
[12] C. Bovet and K. H. Reich, Preprint SI/Int/D1/69-9, CERN, 1969.
[13] G. Holtey, Preprint Lab. 2-DI-PA/Int. 73-3, CERN, 1973.
[14] S. Pegss, Ph.D. Thesis, Cornell, 1981.
[15] L. Burnod and E. D'Amico, IEEE Trans. Nucl. Sci. **NS-30**, No. 4 (1983).
[16] G. Guignard and Y. Marti, Preprint ISR-BOM-TH/81-32, CERN, 1981.
[17] G. Guignard and Y. Marti, Preprint LEP-TH/83-50, CERN, 1983.
[18] E. Bozoki, Preprint PS/PSR/85-57, CERN, 1985.
[19] M. Martini and L. Rinolfi, *Proceedings of EPAC* (Rome, 1988), p. 842.
[20] T. Riselada, Preprint PS/87-90, CERN, 1987.
[21] B. Autin and Y. Marti, Preprint ISR-MA/73-17, CERN, 1973.
[22] Ch. F. Iselin and J. Niederer, Preprint LEP-TH/87-37, CERN, 1987.
[23] Y. Bacconier, Preprint 65-35, CERN, 1965.
[24] G. Guignard, Preprint SI/Int.DI/70-1, CERN, 1970.
[25] G. Guignard, Preprint SI/Int.DI/70-2, CERN, 1970.
[26] G. Guignard, Preprint 70-24, CERN, 1970.
[27] W. Press *et al.*, *Numerical Recipes in C* (Cambridge University Press, 1989).
[28] Y. Chung, G. Decker, and K. Evans, *Proceedings of the IEEE Particle Accelerator Conf.* (Washington, 1993).
[29] D. Diney, Preprint Jul-2406, KFA-Julich, 1980.
[30] Y. Chung *et al.*, Preprint LS-213, ANL, 1992.
[31] D. Dinev and P. Vasilev, Bulg. J. Phys. **12**, No. 5, 480 (1985).
[32] P. K. Simpson, *Artificial Neural Systems* (Pergamon Press, New York, 1989).
[33] E. Bozoki and A. Friedman, *Proceedings of EPAC* (London, 1994), p. 1589.
[34] E. Bozoki and A. Friedman, AIP Conf. Proc. **315** (1994).
[35] A. Zell *et al.* SNNS User Manual, Report 3/13 of the University of Stuttgart, 1993.
[36] R. I. Levine, D. E. Drang, and B. Edelson, *AI and Expert Systems* (McGraw-Hill, New York, 1990).
[37] D. Brandt and A. Verdier, *Proceedings of EPAC* (Rome, 1988), p. 654.
[38] D. Brandt, F. Varlot, and A. Verdier, Part. Accel. **29**, 221 (1990).
[39] C. Bocchetta and A. Wrulich, Nucl. Instrum. Methods A **300**, 223 (1991).
[40] L. H. Yu *et al.* Nucl. Instrum. Methods A **284**, 268 (1989).
[41] Y. Chung, Beam Instrumentation Workshop, Santa Fe, 1993.
[42] D. Dinev, Nucl. Instrum. Methods A **237**, No. 3, 441 (1985).
[43] D. Dinev *et al.*, Zh. Tekh. Fiz. **56**, 1137 (1986) [Sov. Phys. Tech. Phys. **31**, 665 (1986)].
[44] D. Dinev, Preprint Jul-2499, KFA-Julich, 1991.
[45] P. J. Bryant, *CERN Accelerator School* (Gif-sur-Yvette, 1984), p. 358.
[46] J. P. Koutchouk, Preprint LEP-TH/89-2, CERN, 1989.
[47] R. Raya, A. Russell, and C. Aukenbrandt, Nucl. Instrum. Methods A **242**, 15 (1989).
[48] D. Dinev, Bulg. J. Phys. **21**, No. 1 (1994).
[49] V. Paxson, S. Peggs, and L. Schachinger, *Proceedings of EPAC* (Rome, 1988), p. 824.
[50] J. C. Sheppard *et al.* IEEE Trans. Nucl. Sci. **NS-32**, 2180 (1985).
[51] M. Lee *et al.*, IEEE Trans. Nucl. Sci. **NS-34**, 536 (1987).
[52] S. H. Clearwater and M. J. Lee, IEEE Trans. Nucl. Sci. **NS-34**, 532 (1987).
[53] M. J. Lee and S. Kleban, *Proceedings of EPAC* (Rome, 1988), p. 767.
[54] D. P. Weygard, IEEE Trans. Nucl. Sci. **NS-34**, 564 (1987).

This article was published in English in the original Russian journal. It is reproduced here with the stylistic changes by the Translation Editor.