Using kernel-based statistical distance to study the dynamics of charged particle beams in particle-based simulation codes

Chad E. Mitchell, Robert D. Ryne, and Kilean Hwang

Phys. Rev. E **106**, 065302 — Published 8 December 2022

DOI: 10.1103/PhysRevE.106.065302

Using Kernel-Based Statistical Distance to Study the Dynamics of Charged Particle Beams in Particle-Based Simulation Codes

Chad E. Mitchell* and Robert D. Ryne
Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Kilean Hwang
Facility for Rare Isotope Beams, Michigan State University, East Lansing, MI 48824, USA

(Dated: November 18, 2022)

Measures of discrepancy between probability distributions (statistical distance) are widely used in the fields of artificial intelligence and machine learning. We describe how certain measures of statistical distance can be implemented as numerical diagnostics for simulations involving charged-particle beams. Related measures of statistical dependence are also described. The resulting diagnostics provide sensitive measures of dynamical processes important for beams in nonlinear or high-intensity systems, which are otherwise difficult to characterize. The focus is on kernel-based methods such as Maximum Mean Discrepancy, which have a well-developed mathematical foundation and reasonable computational complexity. Several benchmark problems and examples involving intense beams are discussed. While the focus is on charged-particle beams, these methods may also be applied to other many-body systems such as plasmas or gravitational systems.

I. INTRODUCTION

When modeling the dynamics of charged particle beams, the following question often arises. Given two ensembles of simulation particles, how similar are they? In particular, when do two particle ensembles represent the same underlying phase space density? This question is central to validating the random sampling algorithm used for initial beam generation, to comparing particle-based output across multiple simulation codes, to matching the particle beam successfully into a periodic transport system, and for studying the long-time phase space evolution of beams in circular and multi-pass systems.

Two beams are typically compared using their first and second moments, followed by qualitative examination of the beam phase space. For beams at high intensity, in the presence of collective instabilities, or in the presence of highly nonlinear transport, the details of the distribution (including higher-order moments and nonlinear correlations) become increasingly important. Such systems may exhibit filamentation of the beam phase space, developing structure on finer and finer scales, in some cases relaxing to a quasi-stationary state. It is then valuable to have quantitative particle-based diagnostics that can characterize the nonlinear dynamical processes of regular or chaotic mixing and collisionless relaxation [1–6].

These problems can best be addressed by implementing a two-sample measure of statistical distance with well-understood mathematical properties, which can be used to compare particle populations. Statistical distances are widely used in machine learning (ML), information theory, statistics, probability theory, and data mining. Unfortunately, such quantities tend to have high computational complexity. For beam physics applications, the complexity must scale well with the number of particles n (as $O(n^2)$ or better) and with the phase space dimension d (up to at least $d = 6$).

We describe the kernel-based statistical distance known as Maximum Mean Discrepancy (MMD) [7], which has recently had a major impact in the ML community. Broadly speaking, kernel methods allow nonlinear problems involving higher-order statistics to be treated using linear methods, by embedding the set of probability distributions into a reproducing kernel Hilbert space (RKHS) [8, 9]. We do not discuss the RKHS formalism, although many results are most naturally viewed in this context. The use of MMD as a statistical distance leads naturally to a measure of statistical dependence or correlation, known here as the Hilbert Schmidt Correlation (HSCor) [10]. These two diagnostics can provide powerful quantitative tools to study the beam dynamics processes mentioned above.

The structure of this paper is as follows. Section II provides a brief discussion of the beam physics context and a summary of notation. In Section III we review several concepts of statistical distance. Section IV describes the properties of Maximum Mean Discrepancy and its implementation in particle-based tracking codes. Section V describes the properties of the Hilbert Schmidt Correlation and its corresponding implementation. In Section VI, we demonstrate how these tools can be used for beam dynamics applications, using several benchmark problems involving symplectic maps. In Section VII, we apply these tools to examples involving high-intensity beams with self-consistent space charge. Section VIII contains a summary and conclusions. There are three Appendices.

* ChadMitchell@lbl.gov

II. BEAM PHYSICS CONTEXT AND NOTATION

The last two decades have seen increasing progress toward an integrated understanding of the dynamics of many-particle systems with long-range interactions [11–13], their kinetic models [14, 15], and issues related to their asymptotic long-time behavior [16, 17]. Large-scale numerical simulations of such systems require efficient statistical tools to characterize their multiscale behavior. In the related field of fluid turbulence, the turbulent velocity field is characterized using nonlinear statistical tools such as structure functions [19, 20], rates of entropy production [21], intermittency [22], and multifractal methods [18]. In this paper, we focus on a subset of statistical techniques that is especially well-suited to the dynamics of charged particle beams [23, 24].

Consider a beam consisting of N_p identical particles with charge q and mass m subject to electromagnetic self-forces and confined by electromagnetic fields. In an accelerator, such beams are typically localized in space, guided by multipole magnets and accelerated to relativistic speeds by radiofrequency (RF) cavities. Each particle is characterized by its phase space coordinates $X = (x, p_x, y, p_y, z, p_z)$, and the ensemble of particles at time t is described by a probability density f_t on the 6-dimensional phase space, satisfying

$$\int f_t(X)dX = 1, \qquad f_t \geq 0. \tag{1}$$

In a collisionless kinetic description, the evolution of f_t is described by the appropriate Vlasov-Poisson or Vlasov-Maxwell equations, and the condition (1) is preserved during the beam evolution. In storage rings or circular colliders, where beams may circulate for thousands to millions of turns, often one is interested in the long-time behavior of f_t as $t \to \infty$.

Typical quantities of interest include measures of the beam size σ_x, σ_y, σ_z and the beam quality, as characterized by the emittances ϵ_x, ϵ_y, ϵ_z, defined by:

$$\sigma_x^2 = \langle \Delta x^2 \rangle, \qquad \epsilon_x^2 = \langle \Delta x^2 \rangle \langle \Delta p_x^2 \rangle - \langle \Delta x \Delta p_x \rangle^2, \tag{2}$$

where $\Delta x = x - \langle x \rangle$, with corresponding definitions involving y and z. Here $\langle \cdot \rangle$ denotes the expected value with respect to f_t. Minimizing the emittance growth often requires matching the beam to the accelerator focusing fields, finding a t-periodic solution f_t whose period coincides with that of the revolution period.

Such a system is usually modeled by tracking a small number $n << N_p$ of simulated particles along the characteristics of the Vlasov equation, using particle-in-cell or related algorithms. A fundamental problem is to characterize the behavior of f_t using the orbits of the sampled particles, or to compare f_t with a second probability density based on simulated particle samples. This is often done by comparing the first and second-order moments of the phase space coordinates, and quantities such as (2), as estimated from the particle samples. This characterization is sufficient if the beam is well-localized and the forces are sufficiently linear. In cases of nonlinear transport, statistical methods are required that are capable of making additional fine-grained distinctions between distributions.

A. Summary of notation

Because distinct notational conventions appear in the probability, statistics, kinetic theory and statistical physics, machine learning, and beam physics literature, we briefly summarize the conventions used in this paper.

The quantities f, g, and h, with or without subscripts, will denote probability density functions (*e.g.*, beam distribution functions) defined on a fixed phase space M of dimension d. The notation X, X', Y, or Y' will be used to represent points of M, especially when these appear as variables of integration. Integrals are taken over all space, unless otherwise specified. Similarly, the notation X_j, Y_j $(j = 1, 2, \ldots)$ will denote a finite or infinite sequence of points of M, especially when these appear within a summation. The notation k will always denote a kernel function $k : M \times M \to \mathbb{R}$, as defined in Sec. III C.

The notation of Sec. V differs slightly from that of the surrounding sections, in that X and Y, when appearing alone, represent (vector-valued) random variables described by a joint probability density P_{XY}, with marginal densities denoted by P_X and P_Y, respectively.

The problem of interest may now be stated as follows. We wish to compare two probability density functions f and g defined on M. This must be done reliably and efficiently by using only a finite number of points $X_1, \ldots, X_n \in M$ independently sampled from f, and a finite number of points $Y_1, \ldots, Y_m \in M$ independently sampled from g.

III. STATISTICAL DISTANCE

Motivated by the considerations of the previous section, we wish to define a distance ρ between pairs of probability densities f and g on the phase space M, such that ρ satisfies the following natural conditions:

i) non-negativity: $\rho(f, g) \geq 0$,
ii) symmetry: $\rho(f, g) = \rho(g, f)$,
iii) the triangle inequality: $\rho(f, g) \leq \rho(f, h) + \rho(h, g)$ for any probability density h,
iv) identity: $\rho(f, g) = 0$ if and only if $f = g$ (except on a possible set of zero probability).

That is, ρ should define a metric on the set of all probability densities on M [25].

The distance should also capture the concept of relaxation of beams, so that f_t relaxes to f_{eq} in the "coarse-grained" sense if and only if $\rho(f_t, f_{eq}) \to 0$ as $t \to \infty$. The concept of relaxation in the "coarse-grained" sense is well-formalized by the probabilistic concept of weak convergence [26, 27]: we say $f_t \Rightarrow f_{eq}$ if and only if:

$$\lim_{t\to\infty} \int f_t(X)\phi(X)dX = \int f_{eq}(X)\phi(X)dX \qquad (4)$$

for each bounded, continuous function ϕ on the phase space. Informally, (4) states that the ensemble average of each well-behaved observable ϕ must approach the ensemble average of ϕ over f_{eq} as $t \to \infty$. This is a natural concept of convergence to use when characterizing the long-time behavior of beams.

The description of the beam as a dynamically evolving probability density is valid provided that the beam consists of a single particle species, without charge loss. In the presence of charge loss, one is often interested in the behavior of the part of the beam that lies within a bounded subregion of the phase space. The tools described in this paper can also be adapted to treat this case, using a technique to be described in Section VII.

In the following subsections, we briefly describe several indicators of statistical distance that are used in ML and pattern recognition, information science, probability, and statistics. For a general survey, see for example [28].

A. Kullback-Liebler Divergence

The most widely-used statistical distance is the Kullback-Liebler (KL) divergence D_{KL}, given by:

$$D_{KL}(f||g) = \int f(X) \log \frac{f(X)}{g(X)} dX, \qquad (5)$$

which originated in information theory as a measure of the relative entropy of one probability density with respect to another [29, 30]. It has the property that $D_{KL}(f||g) \geq 0$, with equality if and only if $f = g$. (It satisfies metric conditions i and iv above.) However, the integral in (5) is not defined for all pairs of probability densities f and g, and D_{KL} fails to satisfy conditions ii and iii, so it is not a metric in the above sense. A symmetrized modification of KL divergence is often used [30], although this quantity does not satisfy condition iii.

Note that D_{KL} is invariant under any symplectic time-evolution map $\mathcal{M}_t$, since:

$$\begin{aligned} D_{KL}(f_t||g_t) &= \int f_t(X) \log \frac{f_t(X)}{g_t(X)} dX \qquad (6) \\ &= \int f_0(\mathcal{M}_t^{-1}(X)) \log \frac{f_0(\mathcal{M}_t^{-1}(X))}{g_0(\mathcal{M}_t^{-1}(X))} dX \\ &= \int f_0(X') \log \frac{f_0(X')}{g_0(X')} dX' = D_{KL}(f_0||g_0), \end{aligned}$$

where we used the fact that $\mathcal{M}_t$ has Jacobian determinant 1. In particular, if a density f_{eq} is invariant under $\mathcal{M}_t$, in the sense that

$$f_{eq}(\mathcal{M}_t^{-1}(X)) = f_{eq}(X), \qquad (7)$$

then $D_{KL}(f_t||f_{eq})$ is independent of t. Thus, D_{KL} cannot capture weak convergence of the form $f_t \Rightarrow f_{eq}$ to an invariant density f_{eq} under a symplectic time-evolution.

The KL divergence has been used in kinetic simulations [31], and it is well-motivated by statistical mechanics considerations. However, typical algorithms for computing D_{KL} require binning the (possibly highly-filamented) distribution functions f and g, which becomes increasingly problematic in a phase space of dimension > 2. Although gridless two-sample estimation algorithms also exist [32], the rate of convergence with particle number can be arbitrarily slow, with a convergence rate that varies with the distribution. This makes the quantity D_{KL} difficult to estimate reliably from samples, especially in spaces of high dimension. For recent work using kernel-based estimators for D_{KL}, see [33].

B. Wasserstein Metric

The p-Wasserstein distance ($p = 1, 2, \ldots$) is defined by:

$$W_p(f, g) = \left(\min_h \int |X - Y|^p h(X, Y) dX dY \right)^{1/p} \qquad (8)$$

where $|X - Y|$ denotes the Euclidean distance between points X and Y. Here the minimum is taken [34] over all joint probability densities h with marginal densities f and g, so that:

$$f(X) = \int h(X, Y) dY, \quad g(Y) = \int h(X, Y) dX. \qquad (9)$$

The distances W_p originated in the theory of optimal transport [35–37], where the case $p = 1$ is also known as the Kantorovich-Rubinstein metric or the Earth Mover's Distance (EMD).

Note that (8) is guaranteed to be finite when the densities f and g both have finite moments of order p. On the set of all such densities, W_p is known to satisfy all the metric conditions (i-iv). It is also known that W_p correctly captures the concept of weak convergence, in the sense that $W_p(f_t, f_{eq}) \to 0$ (as $t \to \infty$) if and only if $f_t \Rightarrow f_{eq}$ and the pth moments of f_t converge to those of f_{eq}.

Due to its desirable geometric properties, the Wasserstein distance has been applied in ML to tasks such as shape matching, image retrieval, graphics, and to the statistical analysis of detector events in high-energy colliders [38–40]. However, the estimation of W_p from sample data requires solving a linear optimization problem with a computational complexity of $O(n^3 \log n)$, where n is the number of samples [41, 42]. Furthermore, the sample

estimate converges to the population value as $O(n^{-1/d})$ for $d > 2$. This makes W_p challenging to use for practical beam dynamics simulations, which typically require $n \geq 10^5$ and $d \geq 4$. Since there are few cases in which (8) can be determined in closed form, algorithms for computing W_p are also difficult to benchmark.

C. Maximum Mean Discrepancy

A *kernel* k is a symmetric, real-valued function defined on pairs of phase space points that is positive definite, in the sense that:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j k(X_i, X_j) \geq 0, \tag{10}$$

for each $N = 1, 2, \ldots$, each finite set of points $X_1, \ldots, X_N$ and real numbers $c_1, \ldots, c_N$. Typical examples are provided in Appendix A.

To each kernel k is associated a Hilbert space (RKHS) consisting of real-valued functions on the phase space. The Maximum Mean Discrepancy (MMD) between two probability densities f and g is then defined by [7]:

$$\gamma_k(f, g) = \max_{\phi} \left| \int f(X)\phi(X)dX - \int g(X)\phi(X)dX \right| \tag{11}$$

where the maximum is taken over all functions ϕ in the RKHS with $||\phi||_k \leq 1$, $||\cdot||_k$ denotes the Hilbert space norm, and $|\cdot|$ denotes the absolute value.

When k is bounded, the quantity in (11) is defined for all probability densities f and g, and γ_k satisfies the metric conditions (i-iii). Additional restrictions on k are used to ensure that γ_k satisfies condition (iv), and that γ_k captures weak convergence, in the sense previously described. The class of kernels satisfying these restrictions has been extensively studied [43–47], and it includes most of the kernels widely used in ML, including those described in Appendix A.

Due to its well-developed mathematical foundation, its connection to other ML kernel methods such as support vector machines [8], its applicability to domains more general than Euclidean space, and its relative ease of estimation, the distance γ_k has become a powerful tool in statistical two-sample (homogeneity) testing for ML applications. Estimation of (11) from sample data can be achieved using $O(n^2)$ operations, and the sample estimate converges to the population value as $O(n^{-1/2})$, independently of the dimension d. Furthermore, approximations also exist that can be computed with complexity $O(n)$, making γ_k a practical quantity for beam dynamics applications. (Sections IV B-IV C below contain further details.)

IV. PROPERTIES OF MAXIMUM MEAN DISCREPANCY

Given a kernel k, the maximum appearing in (11) can be evaluated exactly by using the properties of its corresponding RKHS. As a result, the MMD between two probability densities f and g can be expressed using the explicit integral formula [43]:

$$\gamma_k(f, g) = \left(\iint k(X, X')\Delta(X)\Delta(X')dXdX' \right)^{1/2}, \tag{12}$$

where $\Delta = f - g$. For certain choices of k, the quantity (12) coincides with other well-known indicators of statistical distance. For example, in the special case that $k(X, X') = |X| + |X'| - |X - X'|$, (12) appears in the statistics literature as the *energy distance* [48, 49].

A. Basic properties

Given any kernel k, one may construct a corresponding kernel k_N by:

$$k_N(X, X') = \frac{k(X, X')}{\sqrt{k(X, X)k(X', X')}}. \tag{13}$$

The condition that k be positive definite (10) then implies that k_N is positive definite with $|k_N| \leq 1$. For simplicity, we will assume that all kernels are so normalized. It then follows from (12) that the distance γ_k is dimensionless with:

$$0 \leq \gamma_k \leq 2. \tag{14}$$

It is natural to choose a kernel that reflects the underlying properties of the domain, so we often consider kernels that are translation-invariant, in the sense that $k(X, X') = k(X + \delta X, X' + \delta X)$ for any phase space displacement δX. A continuous, translation-invariant kernel can be written in terms of its Fourier components as:

$$k(X, X') = \int e^{i(X - X')\cdot\omega} \Lambda(\omega) d\omega. \tag{15}$$

When k is normalized, Λ is a probability density [25] on the space of frequencies $\omega = (\omega_1, \ldots, \omega_d)$. Using (15) in (12), one finds that:

$$\gamma_k(f, g) = \left(\int |\phi_f(\omega) - \phi_g(\omega)|^2 \Lambda(\omega) d\omega \right)^{1/2}, \tag{16}$$

where ϕ_f and ϕ_g denote the Fourier transforms of the densities f and g, normalized so that:

$$\phi_f(\omega) = \int e^{iX\cdot\omega} f(X) dX. \tag{17}$$

When the probability density Λ is also an integrable function that is strictly positive everywhere, it is possible to

prove that (16) satisfies the metric conditions (i-iv) and correctly reflects the weak convergence of probability distributions, as previously described.

One additional property of γ_k is also useful. If $\{e_l : l = 1, 2, \ldots\}$ denotes an orthonormal basis for the RKHS associated with the kernel k, then we may define a complete sequence of beam "moments" m_l by:

$$m_l(f) = \int e_l(X) f(X) dX \quad (l = 1, 2, \ldots). \tag{18}$$

It follows from (12) that the MMD between two distributions f and g may be written in terms of these moments as:

$$\gamma_k(f, g) = \sqrt{\sum_{l=1}^{\infty} |m_l(f) - m_l(g)|^2}. \tag{19}$$

In particular, for every $l = 1, 2, \ldots$ we have:

$$|m_l(f) - m_l(g)| \leq \gamma_k(f, g). \tag{20}$$

Thus when γ_k is small, all of the moments m_l of the distributions f and g must nearly coincide. (An example is provided in Appendix A.)

B. Sample estimate

A direct estimate of (12) from particle (sample) data is given by [7]:

$$\gamma_k^2(f, g) = \frac{1}{m^2} \sum_{i,j=1}^{m} k(X_i, X_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(X_i, Y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(Y_i, Y_j), \tag{21}$$

where the m particle phase space coordinates $\{X_j\}_{j=1}^m$ are independently sampled from the distribution f, and the n particle phase space coordinates $\{Y_j\}_{j=1}^n$ are independently sampled from the distribution g. Note that we allow $m \neq n$.

An alternative grouping of the sum (21) that yields superior numerical precision in practice is given by:

$$\gamma_k^2(f, g) = \sum_{i,j=1}^{m+n} c_i c_j k(\hat{X}_i, \hat{X}_j), \tag{22}$$

where

$$c_j = \begin{cases} \frac{1}{m}, & 1 \leq j \leq m \\ -\frac{1}{n}, & m+1 \leq j \leq m+n \end{cases} \tag{23}$$

and $\{\hat{X}_j\}_{j=1}^{m+n}$ contains the phase space coordinates independently sampled from the distribution f, followed by the phase space coordinates independently sampled from g, so that:

$$\hat{X}_j = \begin{cases} X_j, & 1 \leq j \leq m \\ Y_{j-m}, & m+1 \leq j \leq m+n \end{cases}. \tag{24}$$

The use of (22) avoids a loss of precision that can occur in the evaluation of (21) due to the cancellation of large terms. In the form (22), it is clear from (10) that the estimate satisfies $\gamma_k^2 \geq 0$, and that the computational complexity is $O((m+n)^2)$.

In the special case that the kernel k is translation-invariant, the complexity can be reduced by using the spectral representation (16) to approximate γ_k as the sum [50, 51]:

$$\gamma_k^2(f, g) = \frac{1}{L} \sum_{l=1}^{L} \left| \sum_{j=1}^{m+n} c_j e^{i\omega_l \cdot \hat{X}_j} \right|^2 \tag{25}$$

where the $\{\omega_l\}_{l=1}^L$ denote L frequency vectors that are independently sampled from the spectral probablity density Λ, and the quantities $\{c_j\}_{j=1}^{m+n}$ and $\{\hat{X}_j\}_{j=1}^{m+n}$ are given by (23-24). The complexity of (25) is $O(L(m+n))$, where in practice it is sufficient to use $L \ll m+n$. This results in significant speed-up when translation-invariant kernels are used to compute γ_k.

C. Statistical error

It is shown in [41, 42] that the sample estimate (21) converges to the population value (12) as $O(m^{-1/2} + n^{-1/2})$ when $m, n \to \infty$, independently of the dimension of the underlying space. A similar result [50, 51] may be obtained for the estimate (25).

In practice, the sample estimate in (21-22) or (25) is used to test the hypothesis that two distributions are distinct, $f \neq g$. To set a confidence threshold for this test, one must know the statistical distribution of (21) under the null hypothesis that $f = g$. This problem has been treated in detail [7, 48]. For our purposes, it is enough to know that the root-mean square (rms) value of γ_k under the hypothesis that $f = g$ is given by taking the expected value of (21), which is given by:

$$\gamma_k^{\text{noise}} = \mathrm{E}[\gamma_k^2]^{1/2} = \sqrt{\frac{m+n}{mn}} \left(1 - ||f||_k^2\right)^{1/2}, \tag{26}$$

where the notation $||\cdot||_k$ denotes

$$||f||_k^2 = \iint k(X, X') f(X) f(X') dX dX'. \tag{27}$$

An identical result is obtained by using (25). Thus (26) represents a statistical noise level for γ_k that is associated with the use of a finite number of particles, and when $m = n$, we see that $\gamma_k^{\text{noise}} \propto 1/\sqrt{n}$. The probability P

that γ_k exceeds a threshold value $\tau > 0$ is then bounded above by Markov's inequality, which states that for any $\tau > 0$:

$$P(\gamma_k > \tau) \leq \left(\frac{\gamma_k^{\text{noise}}}{\tau} \right)^2. \tag{28}$$

It follows that (given the null hypothesis) large deviations above the noise floor γ_k^{noise} are rare. A more detailed investigation [7] reveals that the bound in (28) is loose—that is, the probability on the left-hand side in (28) is much smaller and decays more quickly with τ than (28) alone would suggest.

A discussion of statistical error in the more general case (when one may have $f \neq g$) is provided in Appendix C.

D. Numerical implementation

The expressions in (22) and (25) are straightforward to implement in a parallel particle-based beam dynamics simulation code. The resulting numerical diagnostic, which we denote by MMD, may be used to compare an evolving particle distribution f_t with itself after successive t-intervals Δt (*e.g.*, lattice periods), or to compare the evolving particle distribution against a fixed reference distribution (usually an initial or predicted final distribution). The frequency samples $\{\omega_l\}_{l=1}^L$ in (25) may be generated once and stored at initialization of the simulation, or the samples may be drawn independently at each evaluation of the MMD. The latter is the approach favored in the literature.

An algorithm can also be implemented to estimate γ_k using the representation given in (19). When the basis functions $\{e_l\}_{l=1}^\infty$ are known, it is straightforward to estimate the moments (18) from particle samples. However, the number of basis functions required to obtain convergence of the sum (19) grows rapidly with the phase space dimension d. For a Gaussian kernel with $d \geq 2$, we found that this algorithm was outperformed by the spectral algorithm (25) for all examples tested.

A remark about the choice of kernel k is in order. For simplicity, we use a Gaussian kernel (69) along each unbounded phase space coordinate. The kernel width σ is chosen to coincide with a typical rms beam size. In some cases, it is natural to consider periodic domains (for example, if one models a longitudinal beam slice with periodic boundary conditions). Along a phase space coordinate that is naturally periodic, we use a Poisson kernel (73) with the appropriate periodicity. For dynamical problems, it is important that the kernel remain fixed throughout the simulation.

V. HILBERT SCHMIDT CORRELATION

Recall that two random variables X and Y described by a joint probability density P_{XY} are said to be independent when $P_{XY} = P_X P_Y$, where P_X and P_Y are the marginal densities given by:

$$P_X(X) = \int P_{XY}(X,Y)dY, \tag{29a}$$

$$P_Y(Y) = \int P_{XY}(X,Y)dX. \tag{29b}$$

Given a metric ρ on the set of such joint probability densities, a natural measure of deviation from independence between X and Y is given by the distance $\rho(P_{XY}, P_X P_Y)$. This motivates the following definitions.

A. Definition and properties

For simplicity, we assume that X and Y take their values in the same space $\mathbb{R}^d$, on which a kernel k is defined. We define a new kernel κ on $\mathbb{R}^d \times \mathbb{R}^d$ by:

$$\kappa((X,Y),(X',Y')) = k(X,X')k(Y,Y'). \tag{30}$$

It follows that κ is symmetric and positive definite (10).

The Hilbert Schmidt correlation $\mathcal{R}_k$ between X and Y is then defined by:

$$\mathcal{R}_k^2(X,Y) = \frac{\gamma_\kappa^2(P_{XY}, P_X P_Y)}{\gamma_\kappa(P_{XX}, P_X P_X)\gamma_\kappa(P_{YY}, P_Y P_Y)}. \tag{31}$$

The quantity in the numerator is known as the Hilbert Schmidt Independence Criterion (HSIC) [10, 52, 53]. The normalizing factor in the denominator appears in [48], and is designed to ensure that $\mathcal{R}_k \leq 1$. The joint densities P_{XX} and P_{YY} represent the limiting case of perfect correlation when $X = Y$, namely:

$$\begin{aligned} P_{XX}(X,Y) &= P_X(X)\delta(Y-X), \\ P_{YY}(X,Y) &= P_Y(Y)\delta(X-Y). \end{aligned} \tag{32}$$

Here δ denotes the Dirac delta.

It may be shown that the quantity in (31) satisfies:

$$\begin{aligned} &0 \leq \mathcal{R}_k \leq 1, \\ &\mathcal{R}_k = 0 \text{ if and only if } X \text{ and } Y \text{ are independent}, \\ &\mathcal{R}_k = 1 \text{ if } X \text{ and } Y \text{ are identical}. \end{aligned} \tag{33}$$

It follows that $\mathcal{R}_k$ provides a natural measure of (possibly nonlinear) correlation between X and Y. The special case when $k(X,X') = |X| + |X'| - |X - X'|$ is known in the statistics literature as the *distance correlation* (dCor) [48, 49, 54, 55].

When the kernel k is translation-invariant, we may use the representation (16) to write the numerator of (31) as:

$$\gamma_\kappa^2(P_{XY}, P_X P_Y) = \int \left| \phi_{XY}(\omega,\omega') - \phi_X(\omega)\phi_Y(\omega') \right|^2 \Lambda(\omega)\Lambda(\omega')d\omega d\omega', \tag{34}$$

where Λ is the spectral density of k defined in (15), and

$$\phi_{XY}(\omega,\omega') = \int e^{i(\omega\cdot X + \omega'\cdot Y)} P_{XY}(X,Y)dXdY, \tag{35a}$$

$$\phi_X(\omega) = \phi_{XY}(\omega,0), \qquad \phi_Y(\omega') = \phi_{XY}(0,\omega'). \tag{35b}$$

Corresponding expressions for the factors in the denominator of (31) are obtained from (34) by taking $Y \mapsto X$ or $X \mapsto Y$ as appropriate, and using (32) to write:

$$\begin{aligned}\phi_{XX}(\omega,\omega') &= \int e^{i(\omega\cdot X+\omega'\cdot Y)} P_{XX}(X,Y)dXdY \\ &= \phi_X(\omega+\omega'), \qquad (36a)\\ \phi_{YY}(\omega,\omega') &= \int e^{i(\omega\cdot X+\omega'\cdot Y)} P_{YY}(X,Y)dXdY \\ &= \phi_Y(\omega+\omega'). \qquad (36b)\end{aligned}$$

B. Sample estimate

An estimate of the numerator of (31) from particle (sample) data is given by:

$$\begin{aligned}\gamma_\kappa^2(P_{XY},P_XP_Y) = &\frac{1}{m^2}\sum_{i,j=1}^{m} k(X_i,X_j)k(Y_i,Y_j) \\ &+\frac{1}{m^4}\sum_{i,j,q,r=1}^{m} k(X_i,X_j)k(Y_q,Y_r) \\ &-\frac{2}{m^3}\sum_{i,j,q=1}^{m} k(X_i,X_j)k(Y_i,Y_q) \qquad (37)\end{aligned}$$

where the m pairs $\{(X_j,Y_j)\}_{j=1}^m$ are independently sampled from the density P_{XY}. The corresponding expressions appearing in the denominator of (31) are obtained from (37) by taking $Y_j \mapsto X_j$ and $X_j \mapsto Y_j$, respectively. It can be shown that (37) can be computed with $O(m^2)$ complexity by introducing $O(m)$ storage.

In the special case that the kernel k is translation invariant, this complexity can be further reduced by working in frequency space using (34). An efficient estimate is given by:

$$\begin{aligned}\gamma_\kappa^2(P_{XY},P_XP_Y) = \frac{1}{L}\sum_{k=1}^{L}\Bigg|&\frac{1}{m}\sum_{j=1}^{m} e^{i\omega_{2k-1}\cdot X_j}e^{-i\omega_{2k}\cdot Y_j} \\ &-\left(\frac{1}{m}\sum_{j=1}^{m} e^{i\omega_{2k-1}\cdot X_j}\right)\left(\frac{1}{m}\sum_{j=1}^{m} e^{-i\omega_{2k}\cdot Y_j}\right)\Bigg|^2, \qquad (38)\end{aligned}$$

where $\{\omega_k\}_{k=1}^{2L}$ denote $2L$ frequency vectors that are independently sampled from the spectral probability density Λ associated with k. Note that the complexity of (38) is given by $O(mL)$. (This is a variant of the random Fourier features estimate appearing in [51, 56].)

C. Statistical error

In practice, the sample estimate in (37) or (38) is used to test the hypothesis that two random variables X and Y are independent. To set a confidence threshold for this test, one needs to know the statistical distribution of (37) under the null hypothesis that $P_{XY} = P_XP_Y$. This problem has been treated in detail [10, 48, 52, 53]. For our purposes, it is enough to know the rms value of $\mathcal{R}_k$ under the hypothesis that X and Y are independent. This is given by taking the expected value of (37), which yields:

$$\begin{aligned}&\mathrm{E}[\gamma_k^2(P_{XY},P_XP_Y)] \\ &\quad= \frac{(m-1)}{m^2}\left(1-||P_X||_k^2\right)\left(1-||P_Y||_k^2\right), \qquad (39)\end{aligned}$$

where $||\cdot||_k$ has the same meaning as in (27). Thus we have, to leading order in $1/m$:

$$\mathcal{R}_k^{\mathrm{noise}} = \frac{1}{\sqrt{m}}\frac{\left(1-||P_X||_k^2\right)^{1/2}\left(1-||P_Y||_k^2\right)^{1/2}}{\gamma_\kappa(P_{XX},P_XP_X)\gamma_\kappa(P_{YY},P_YP_Y)}. \qquad (40)$$

Note that (40) is fully determined by the marginal densities P_X and P_Y through (32).

Given the null hypothesis, an inequality corresponding to (28) holds after γ_k is replaced by $\mathcal{R}_k$, indicating that large deviations above the noise floor value $\mathcal{R}_k^{\mathrm{noise}}$ are rare.

D. Numerical Implementation

In practice, the random variables X and Y described above may represent two phase space coordinates within a single beam (*e.g.*, $X = z$ and $Y = p_z$) or two d-tuples of phase space coordinates (*e.g.*, $X = (x, p_x)$ and $Y = (y, p_y)$). In this case, computation of $\mathcal{R}_k(X,Y)$ using (37) or (38) returns a measure of correlation between phase space coordinates within the beam.

Alternatively, let $X(0)$ denote the vector of phase space coordinates for a particle within the beam at initial time (or lattice location) $t = 0$, and let $X(t)$ denote the vector of phase space coordinates for the same particle at a later time t. Then $\mathcal{R}_k(X(0),X(t))$ measures the correlation of a particle's coordinates at time t with the particle's initial coordinates, and this quantity will be denoted $\mathcal{R}_k(t)$ for simplicity. The dynamical evolution of this quantity is intimately related to mixing. (See Appendix B.)

Given a beam consisting of m particles, numerical evaluation of $\mathcal{R}_k(t)$ requires that each particle be assigned a unique index j, so that one may construct the set of particle pairs $(X(0)_j, X(t)_j)$, $j = 1,\ldots,m$ at each desired evaluation time t. In particular, the arrays containing the initial and final phase space coordinates of particle j must be stored on the same processor, which requires appropriate bookkeeping and possible communication.

VI. APPLICATIONS TO IDEALIZED AND EXACTLY-SOLVABLE MODELS

In this section, we illustrate several applications of the above tools to dynamical problems involving beams: 1)

to compare two beams for benchmarking and quantifying beam mismatch, 2) to detect nonlinear phase space correlations and coupling between phase planes, 3) to verify numerically that a beam that is generated from a well-matched distribution remains stationary, 4) to study the relaxation of a non-equilibrium beam to a stationary state, and 5) to measure the rate of chaotic mixing and decay of correlations within a beam during its evolution.

To aid in benchmarking the diagnostics MMD (γ_k) and HSCor ($\mathcal{R}_k$), idealized distributions and exactly-solvable models involving maps are used. The next section will discuss realistic examples involving high-intensity beams.

A. Distribution comparison and mismatch

Let f and f' denote two centered Gaussian distributions with covariance matrices Σ and Σ', respectively. To determine the distance between these distributions, we consider an arbitrary Gaussian kernel k of the form:

$$k(X, X') = \exp\left(-\frac{1}{2}(X - X')^T S(X - X')\right), \quad (41)$$

where S is any symmetric, positive definite matrix. Such a matrix may always be decomposed as $S = A^T A$ for some matrix A. The MMD between f and f' may be obtained using (12) as:

$$\gamma_k^2(f, f') = \det(I + 2\Sigma_N)^{-1/2} + \det(I + 2\Sigma'_N)^{-1/2} - 2\det(I + \Sigma_N + \Sigma'_N)^{-1/2}, \quad (42)$$

where I is the identity matrix, and the normalized covariance matrices are:

$$\Sigma_N = A\Sigma A^T, \qquad \Sigma'_N = A\Sigma' A^T. \quad (43)$$

This yields a large class of examples for benchmarking the computation of MMD in any dimension.

As a special case of (42), consider two Gaussian beams f and f' described on a 2D phase space (x, p_x) with identical emittance ϵ, with covariance matrices:

$$\Sigma = \epsilon \begin{pmatrix} \beta & -\alpha \\ -\alpha & \gamma \end{pmatrix}, \qquad \Sigma' = \epsilon \begin{pmatrix} \beta' & -\alpha' \\ -\alpha' & \gamma' \end{pmatrix}. \quad (44)$$

Taking $S = \Sigma^{-1}$ in (41) and computing the MMD between f and f' using (42) gives:

$$\gamma_k^2(f, f') = \frac{1}{3} - \frac{1}{\sqrt{5 + 4\zeta}}. \quad (45)$$

Here the result is expressed in terms of the linear beam mismatch parameter ζ, given by:

$$\zeta = \frac{1}{2}\left(\beta\gamma' - 2\alpha\alpha' + \gamma\beta'\right), \quad \zeta \geq 1. \quad (46)$$

The same result is obtained by taking $S = (\Sigma')^{-1}$ in (41). Note that (45) vanishes when $\zeta = 1$, and increases monotonically with increasing mismatch ζ. Thus, when the MMD can be expressed in terms of the linear mismatch, the result behaves as expected.

In addition to detecting differences based on the second beam moments, the MMD detects differences in the details of two distributions. For a 4D example relevant to charged-particle beams, consider a K-V distribution f_{KV} and a (4D) Gaussian distribution f_G with the same second moments, described by a 4×4 covariance matrix Σ in the variables $X = (x, p_x, y, p_y)$. By choosing the kernel (41) with $S = \Sigma^{-1}$, we may take without loss of generality $\Sigma = I_{4\times4}$, the 4×4 identity matrix. Then we have:

$$f_{KV}(x, p_x, y, p_y) = \frac{1}{2\pi^2}\delta(4 - ||X||^2), \quad (47a)$$

$$f_G(x, p_x, y, p_y) = \frac{1}{(2\pi)^2} e^{-||X||^2/2}, \quad (47b)$$

where $||X|| = (x^2 + p_x^2 + y^2 + p_y^2)^{1/2}$ and δ denotes a Dirac delta. Taking the Fourier transforms of (47) according to (17) and evaluating the integral in (16) gives the exact result:

$$\gamma_k(f_{KV}, f_G) = \sqrt{\frac{1}{9} - \frac{1}{2e} + \frac{I_2(4) + I_3(4)}{2e^4}}, \quad (48)$$

where I_n denotes the modified Bessel function of order n. This result corresponds to the numerical value $\gamma_k(f_{KV}, f_G) \approx 0.12863$. Figure 1 illustrates the numerical error associated with the estimation of this quantity using (25), for varying number of particles $n = m$ and number of frequency samples L. See Appendix C for further discussion of the numerical error.

B. Detecting phase space correlations

For a Gaussian distribution of any dimension, one may detect linear correlations among the various degrees of freedom by using (42) in (31). As an example, consider a Gaussian distribution on a 2D phase space (q, p) with the covariance matrix:

$$\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \quad f(q, p) = \frac{1}{2\pi a} e^{-(q^2 + p^2 - 2rqp)/2a^2}, \quad (49)$$

where $a = \sqrt{1 - r^2}$ and $-1 \leq r \leq 1$. Using a Gaussian kernel of unit width, the correlation $\mathcal{R}_k$ between the variables q and p is given by:

$$\mathcal{R}_k^2 = \frac{g(r)}{g(1)}, \quad g(r) = \frac{1}{3} + \frac{1}{\sqrt{9 - 4r^2}} - \frac{2}{\sqrt{9 - r^2}}. \quad (50)$$

This result is expressed in terms of the standard linear correlation coefficient r. Note that $\mathcal{R}_k$ increases monotonically from 0 to 1 as $|r|$ increases from 0 to 1. Thus, when the HSCor can be expressed in terms of the linear correlation coefficient, the results behaves as expected.

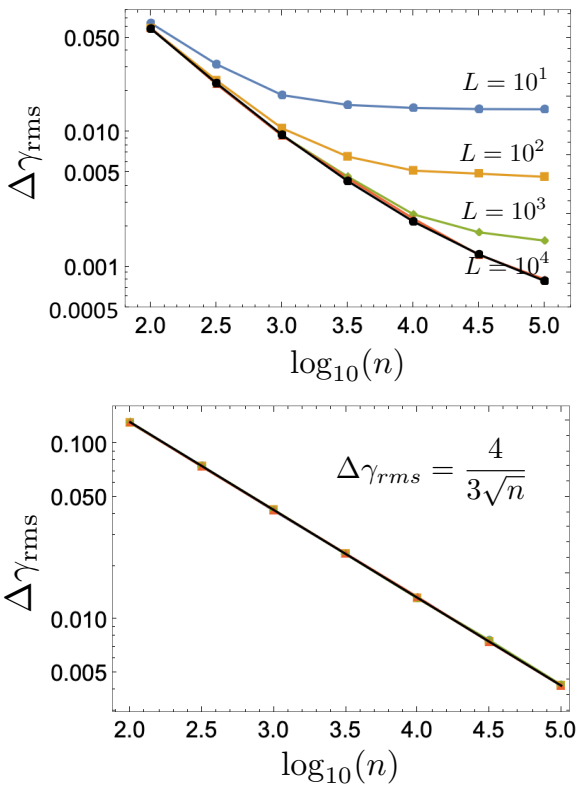

FIG. 1. (Upper) Difference between the result $\gamma_k(f_{KV}, f_G)$ given in (48) and its numerical estimate obtained using (25) for varying number of particles $n = m$ and number of frequency samples L. Statistical averaging was performed using 1000 distinct realizations of the sampled distributions f_{KV}, f_G, and Λ. Curves for increasing values of L approach the curve obtained using the estimate based on (22), shown in black. (Lower) The same quantities, shown for the case of two (4D) Gaussian distributions with the same second moments. In this case, the rms error $\Delta\gamma_{\rm rms}$ is independent of the number of frequency samples L and is equal to $\gamma_k^{\rm noise}$ (26).

The quantity $\mathcal{R}_k$ is useful for detecting nonlinear correlations, even when the exact structure of the correlation is unknown. For example, a quadratic correlation between longitudinal position z and momentum p_z can appear after transporting a charged-particle beam through a radiofrequency (RF) accelerating cavity at the appropriate phase. Consider the case of a Gaussian beam with such a quadratic correlation in the longitudinal phase space (z, p_z):

$$f(z, p_z) = \frac{1}{2\pi\sigma_z\sigma_p} e^{-z^2/2\sigma_z^2} e^{-(p_z + az^2)^2/2\sigma_p^2}. \qquad (51)$$

Here σ_z denotes the longitudinal beam size, σ_p denotes the (uncorrelated) momentum spread, and $a \neq 0$. The linear correlation between z and p_z in (51) vanishes, since one may verify that $\langle z p_z \rangle = 0$.

Using the dimensionless variables $\bar{z} = z/\sigma_z$ and $\bar{p}_z = p_z/(a\sigma_z^2)$, and choosing a Gaussian kernel of width 1,

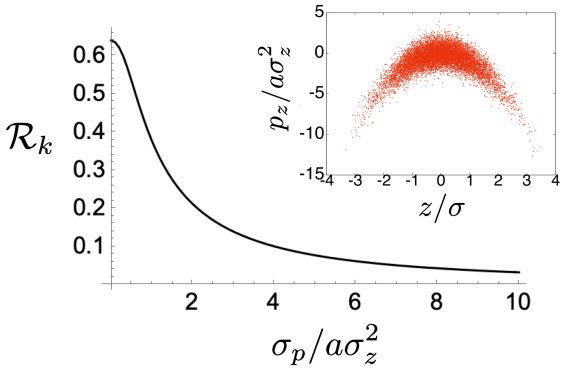

FIG. 2. Hilbert-Schmidt correlation $\mathcal{R}_k$ between variables z and p_z in the longitudinal phase space of a bunch with a quadratic correlation (51). (Inset) Sampled particles (10^4) for the case $\sigma_p/a\sigma_z^2 = 1$, yielding the computed value $\mathcal{R}_k \approx 0.375$.

one may evaluate the correlation $\mathcal{R}_k$ between $\bar{z}$ and $\bar{p}_z$. This may be achieved by using the definition (31) with random variables $X = \bar{z}$, $Y = \bar{p}_z$, and evaluating the integrals in (34-35) numerically. The result is shown in Fig. 2 as a function of the dimensionless parameter $\bar{\sigma}_p = \sigma_p/(a\sigma_z^2)$. The result is independent of the sign of a. We see that $\mathcal{R}_k$ becomes small as the quadratic coefficient a becomes small or as the uncorrrelated momentum spread σ_p becomes large, as one might expect.

C. Testing for stationarity

Often one must characterize the degree to which a given particle distribution remains stationary over many periods of dynamical evolution. This can be quantified by computing the distance $\gamma_k(f_t, f_0)$ between the initial distribution f_0 and the distribution f_t after t periods.

For example, consider the 2D nonlinear symplectic map given by [58]:

$$\begin{pmatrix} q^f \\ p^f \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix}, \quad \phi = \psi + \frac{\alpha}{2}\left(q^2 + p^2\right), \qquad (52)$$

where $\psi > 0$ and $\alpha > 0$ are constants. This may be viewed as a simple model of a betatron phase advance in a single plane that increases linearly with the action $J = (q^2 + p^2)/2$.

A Gaussian distribution of the form:

$$f(q, p) = \frac{1}{2\pi\epsilon_0} e^{-(q^2+p^2)/2\epsilon_0} \qquad (53)$$

is an explicit function of the action J, and is therefore invariant under the map (52). Sampling $n = 10^4$ particles from (53) and tracking them under iterates of the map (52), we compare the distribution at each iteration with the initial distribution. The result is shown in Fig. 3.

The MMD distance $\gamma_k(f_t, f_0)$ to the initial distribution is nonzero after the first iteration, but remains near 10^{-2}

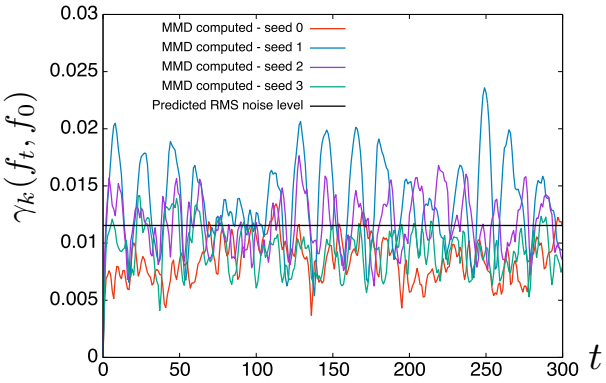

FIG. 3. Dynamics of a 2D beam with $n = 10^4$ particles sampled from a matched Gaussian distribution (53) evolving under iteration of the map (52). The quantity $\gamma_k(f_t, f_0)$ is shown as a function of the iteration number t for 4 distinct random seeds, showing that the distribution remains stationary.

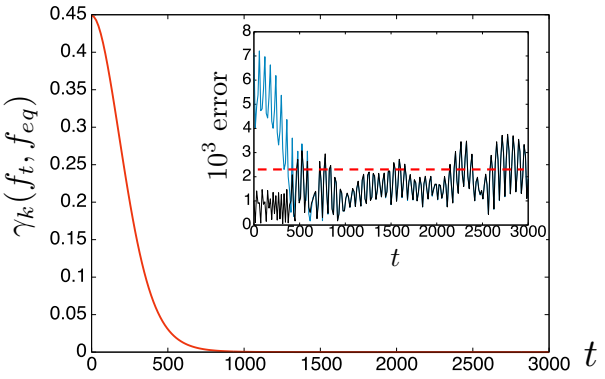

FIG. 4. Evolution of a distribution with $n = 10^5$ particles sampled from (53) with initial offset $q \mapsto q + q_0$ under iteration of the map (52). The distance of the distribution to the predicted equilibrium (54) after t iterates is shown. (Red curve) Analytical prediction (55). (Black, inset) Absolute error in the numerical result obtained using (22). (Blue, inset) Absolute error in the numerical result obtained using (25) with $L = 10^3$. (Red, dashed) Prediction (26) of the numerical noise level evaluated using the distribution f_{eq}.

over the time interval observed. Due to the finite number of particles, the value of $\gamma_k(f_t, f_0)$ experiences statistical fluctuations around the predicted rms value given by (26) (black curve).

D. Relaxation to a stationary state

If the distribution (53) is given an initial centroid offset $q \mapsto q + q_0$ with $q_0 > 0$, the beam will filament [57] and converge weakly to a stationary equilibrium of the form [58]:

$$f_{eq}(q,p) = \frac{1}{2\pi\epsilon_0} e^{-(q^2+p^2+q_0^2)/2\epsilon_0} I_0\left(\frac{q_0}{\epsilon_0}\sqrt{q^2+p^2}\right). \tag{54}$$

See Fig. 1 of [58] for a visual illustration. Using a Gaussian kernel of width σ in expression (16), one may solve exactly for the time evolution of the MMD distance to equilibrium:

$$\gamma_k^2(f_t, f_{eq}) = 2s^2 \sum_{n=1}^{\infty} \nu_n e^{-\nu_n(1+2d^2 s^2 \tau_n^2)} I_n(\nu_n), \tag{55}$$

where I_n denotes the modified Bessel function of order n and

$$\nu_n = \frac{d^2}{1 + d^2 s^2 (1+\tau_n^2)}, \tag{56}$$

is given in terms of the dimensionless parameters:

$$\tau_n = nt\alpha\epsilon_0, \quad d^2 = \frac{q_0^2}{2\epsilon_0}, \quad s = \frac{\sigma}{q_0}. \tag{57}$$

In particular, we see that for large t, (55) converges to zero as $\gamma_k(f_t, f_{eq}) \sim O(1/t^2)$.

Fig. 4 illustrates the result obtained from tracking 10^5 particles sampled from (53) for $\epsilon_0 = 0.01$ with a centroid offset of $q_0 = 0.5$ under the map (52) with $\psi = 0.3$, $\alpha = 0.1$. We use a Gaussian kernel with width parameter $\sigma = 1$ to compare the distribution at each iteration with the stationary distribution (54). The inset shows the difference between the computed value of $\gamma_k(f_t, f_{eq})$ and the analytical prediction for provided in (55). By $t = 500$, the distribution f_t has converged to the distribution f_{eq} to within the resolution set by the particle noise (26). This shows that MMD provides a diagnostic capable of measuring the dynamical relaxation of a beam to a stationary state.

E. Mixing and decay of correlations

Although the term "mixing" is sometimes used to refer to any process involving filamentation and relaxation of the beam in phase space, we distinguish between regular mixing (which is characteristic of nonlinear integrable systems) and mixing in the ergodic sense (which is characteristic of systems with widespread chaos). Here, we refer to the latter meaning of the term, as it is formalized in ergodic theory [59]. (See Appendix B.)

A simple illustration of chaotic mixing behavior is given by the Arnold cat map [60], which is the 2D area-preserving map given by:

$$\begin{pmatrix} q^f \\ p^f \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix} \mod 2\pi, \tag{58}$$

where we assume that q and p each have period 2π [61].

The uniform density:

$$f(q,p) = \frac{1}{(2\pi)^2}, \quad q,p \in [0,2\pi) \tag{59}$$

is invariant under (58) since the map is area-preserving. Figures illustrating the filamentation and folding induced by the cat map may be found in many places [62].

The Hilbert Schmidt correlation provides a quantitative measure of mixing and the resulting decay of correlations over time. To illustrate this, we sample $n = 10^5$ initial conditions from the distribution (59) and track these under the map (58). After each iteration, we compute the value $\mathcal{R}_k(X(t), X(0))$, where $X(0)$ is the random variable denoting a particle's initial phase space coordinates, and $X(t)$ denotes the particle's phase space coordinates after t iterations of the map. For simplicity, we denote this quantity by $\mathcal{R}_k(t)$. In a periodic domain, it is appropriate to use a kernel that reflects the underlying domain periodicity. Using a Poisson kernel with parameter σ in each dimension (Appendix A), we may compute the value of $\mathcal{R}_k(t)$ explicitly. The value after t iterations is given exactly by the sum:

$$\mathcal{R}_k(t) = \left(\frac{1-\sigma^2}{2\sigma}\right)\left[\sum_{\substack{n_1,n_2=-\infty \\ n_1\neq 0, n_2\neq 0}}^{\infty} \sigma^{C(n_1,n_2,t)}\right]^{1/2}, \tag{60}$$

where the exponent is ($t = 1, 2, 3, \ldots$):

$$\begin{aligned} C(n_1,n_2,t) &= |n_1| + |n_2| \\ &+ |F_{2t+1}n_1 + F_{2t}n_2| + |F_{2t}n_1 + F_{2t-1}n_2| \end{aligned} \tag{61}$$

given in terms of the usual Fibonacci sequence:

$$F_1 = F_2 = 1, \quad F_t = F_{t-1} + F_{t-2}. \tag{62}$$

Fig. 5 shows the decay of the quantity $\mathcal{R}_k(t)$ as a function of the iteration number t for the case $\sigma = 1/2$. Because the prediction (60) is only defined for nonnegative integer values of t, the red curve shown connects these values using smooth interpolation. Note that mixing for the map (58) is very rapid. After only 4-5 iterations, the correlations between the initial and final phase space coordinates are at or below the level expected due to numerical particle noise, as given by (40) and indicated by the black line.

To illustrate the dynamical decay of correlations in more detail, Fig. 6 shows the initial phase space coordinates $(q(0), p(0))$ versus the coordinate $q(t)$ at several times $t > 0$, for each of the sampled particles used to generate Fig. 5. Nonlinear correlations are visible, and these correlations occur at a scale length that decreases with increasing t, until by $t = 10$ no visible correlations remain, consistent with the behavior of $\mathcal{R}_k(t)$ in Fig. 5. (The plots are shown in 3D because the correlations between initial and final coordinates are not visible using 2D projections.)

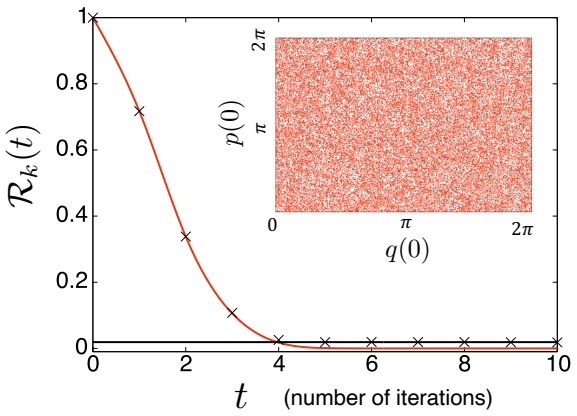

FIG. 5. Dynamics of a beam with $n = 10^5$ particles sampled from the density (59) evolving under iteration of the map (58). The quantity $\mathcal{R}_k(t)$ is shown as a function of the iteration number t, illustrating the decay of correlations due to mixing. The red curve shows the prediction (60), the black points are the results of simulation, and the black curve denotes the expected rms value due to noise (40). (Inset) Plot of initial particle coordinates $(q(0), p(0))$ sampled from the density (59).

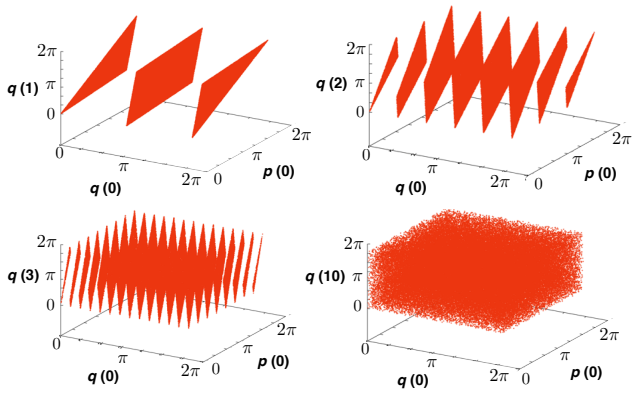

FIG. 6. Initial phase space coordinates $(q(0), p(0))$ versus the final coordinate $q(t)$ at $t = 1, 2, 3, 10$ for the particles used to generate Fig. 5, showing the visible correlations that are quantified by $\mathcal{R}_k(t)$ and their evolution over time.

VII. APPLICATION TO HIGH-INTENSITY BEAMS

A. Beam in a constant focusing channel

As our first example including self-field effects (space charge), we consider an unbunched intense beam in a constant focusing channel that is initialized in a stationary thermal equilibrium, so the 4D beam distribution takes the form:

$$f_0(x, p_x, y, p_y) \propto e^{-H(x,p_x,y,p_y)/kT}, \tag{63}$$

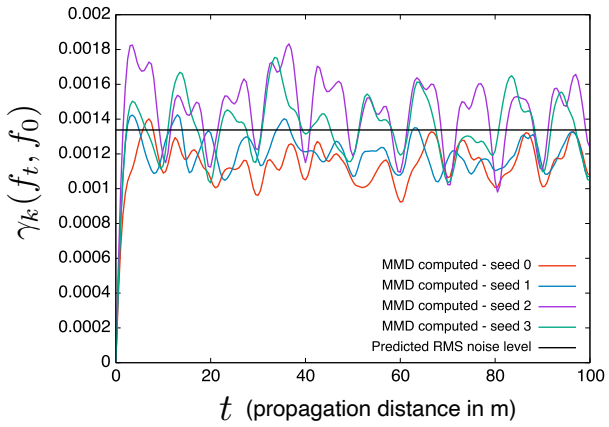

FIG. 7. Test of stationarity for an unbunched (4D) beam with $n = 10^6$ particles sampled from a thermal equilibrium distribution (63) propagating in a linear constant focusing channel. The quantity $\gamma_k(f_t, f_0)$ is shown as a function of propagation distance t for 4 distinct random seeds.

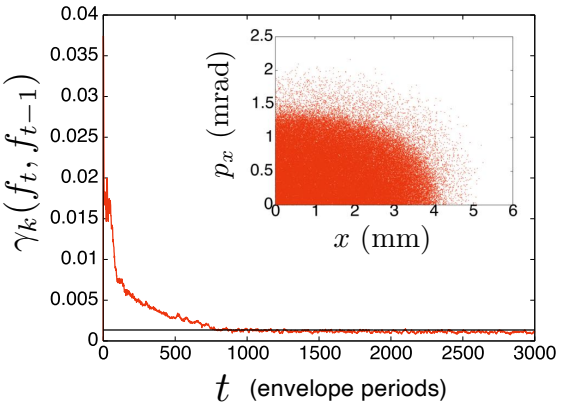

FIG. 8. Collisionless relaxation of an unbunched (4D) beam with $n = 10^6$ particles sampled from the distribution (66) in a linear constant focusing channel (64), where t denotes the number of linearized envelope periods. The MMD between the distribution on successive periods $\gamma_k(f_t, f_{t-1})$ decays to the level of noise as the beam relaxes to a stationary state. (Inset) The final particle distribution (one quadrant).

where H denotes the self-consistent Hamiltonian:

$$H = \frac{1}{2}(p_x^2 + p_y^2) + \frac{1}{2}\Omega^2(x^2 + y^2) + \frac{q\phi(x,y)}{\beta^2\gamma^3 mc^2}, \qquad (64)$$

and ϕ is a solution of the 2D Poisson equation:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\phi = -\frac{\lambda}{\epsilon_0}\int f_0(x, p_x, y, p_y)dp_x dp_y. \qquad (65)$$

Here λ denotes the charge per unit length, and ϵ_0 denotes the vacuum permittivity. This model describes an infinitely long monoenergetic beam of uniform line density in z, moving with velocity $\beta = v_z/c$, that is confined using transverse focusing. As usual, $\gamma = 1/\sqrt{1-\beta^2}$.

For simulation, we consider a proton beam with a kinetic energy of 200 MeV and a beam current of 20 A in an external focusing of strength $\Omega = 0.628\ m^{-1}$ (corresponding to a 2.7 T solenoid field). The temperature kT is chosen to yield the initial emittance $\epsilon_{x,rms} = \epsilon_{y,rms} = 1.24\ \mu$m. The tune depression due to space charge is then given by $\nu/\nu_0 \approx 0.55$. See [23] for additional details regarding the physical model.

Fig. 7 illustrates the MMD distance between the initial distribution and the distribution at time t, for four distinct random seeds. To compute (25), a Gaussian kernel was used. The kernel width parameter σ associated with each phase space dimension was matched to the corresponding rms width of the distribution (63). For each random seed, self-consistent tracking with transverse (2D) space charge using $n = 10^6$ particles was performed using the symplectic gridless spectral solver described in [63]. Notice that the distribution remains stationary to within the level expected due to particle noise (black line). Compare this dynamical behavior to that shown in Fig. 3.

As a second example, consider a proton beam with the same energy and emittance as above. However, instead of the stationary distribution (63), we use the initial distribution [6]:

$$f_0(x, p_x, y, p_y) = \frac{1}{\pi^2 r_m^2 p_m^2}\Theta(r_m - r)\Theta(p_m - p) \qquad (66)$$

where $r = \sqrt{x^2 + y^2}$, $p = \sqrt{p_x^2 + p_y^2}$, and Θ denotes the unit step function. The distribution (66) is not stationary, but to minimize fluctuations of the rms beam size, we match the beam in an rms sense by setting $p_m^2 = \Omega^2 r_m^2 - K_{pv}$, where K_{pv} denotes the generalized beam perveance [23]. In the absence of precise knowledge of the final equilibrium state, we compare the particle distribution at successive time intervals separated by $\Delta t = 6.20$ m, corresponding to the period of linearized envelope oscillations about the equilibrium beam size r_m. The result is shown in Fig. 8. On the time scale shown, the beam appears to undergo relaxation toward a final distribution containing a small but visible low-density halo (inset). This fast relaxation appears to be a property of the collisionless Vlasov-Poisson system, and it is to be distinguished from slow relaxation due to collisional effects [4], which are not included in the numerical model.

An illustration of the time evolution of the particle distribution for this example is provided in Fig. 9. In order to visualize the dynamics of the phase space points, a subset of initial conditions is highlighted in black, and this particle subpopulation is shown at each time t. The phase space region highlighted in black filaments and folds, and at $t = 2000$ this particle subpopulation appears distributed throughout the beam. Nevertheless,

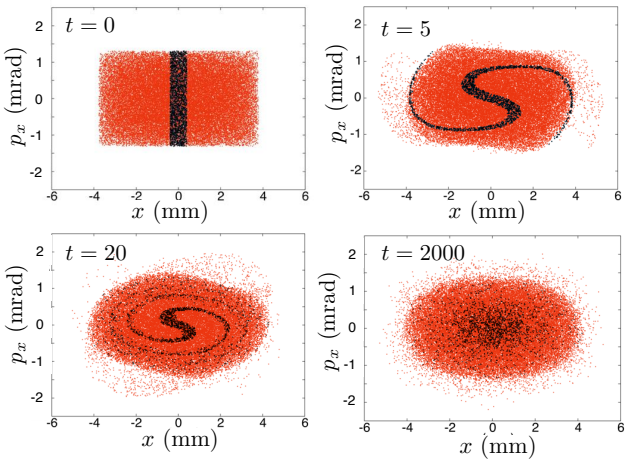

FIG. 9. Projections into the (x, p_x) plane of an initially rms-matched (but non-stationary) distribution (66) evolving in a linear constant focusing channel, shown at four distinct times. The particles shown are those used to produce Fig. 8. (Red) The complete simulated particle distribution. (Black) A subset of initial conditions, illustrating dynamical filamentation of the phase space.

the system described in this section cannot be fully mixing due to the existence of invariants of motion (for example, the angular momentum). Indeed, the computed quantity $\mathcal{R}_k(t)$ converges rapidly to a nonzero value. This behavior of $\mathcal{R}_k(t)$ will be examined using a more complex example in the following section.

B. Beam in a periodic FODO channel

As a typical application to an intense beam in a periodic focusing structure, we consider an alternating gradient quadrupole (FODO) channel. Each period consists of a single FODO cell as shown in Fig. 10. We consider a proton beam with a kinetic energy of 200 MeV and a beam current of 10 A with an initial emittance of $\epsilon_{x,rms} = \epsilon_{y,rms} = 1$ μm (unnormalized). The initial distribution is Gaussian of the form:

$$f_0(x, p_x, y, p_y) = \frac{1}{(2\pi)^2 \epsilon_{x,rms}\epsilon_{y,rms}} e^{-\frac{1}{2}X^T\Sigma^{-1}X}, \quad (67)$$

where $X = (x, p_x, y, p_y)$ and Σ denotes the covariance matrix. Although the second beam moments are chosen so that the beam is matched in an rms-sense, the distribution is not matched in detail. Fig. 10 shows the evolution of the matched beam envelopes over a single period. The zero-current phase advance is 60.1° per cell, while the 10 A phase advance is 25.9° per cell, so that space charge plays a significant role.

Fig. 11 illustrates the MMD distance between the initial distribution and the distribution after t periods (blue), together with the MMD distance between successive periods (red). To compute (25), a Gaussian kernel

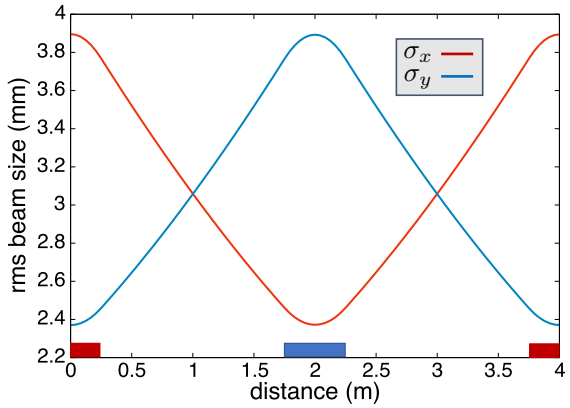

FIG. 10. Matched (K-V) beam envelopes for a 10 A proton beam at 200 MeV in the FODO cell used in Section VII B. Red rectangle - focusing quadrupole. Blue rectangle - defocusing quadrupole.

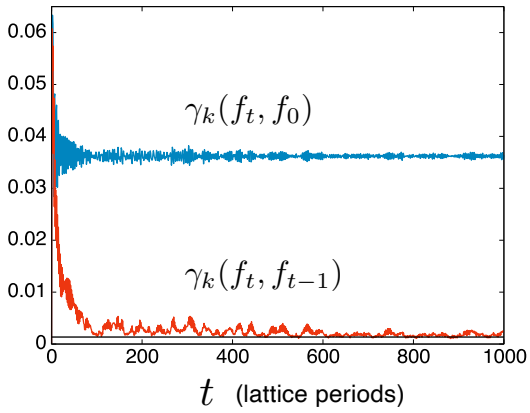

FIG. 11. Dynamics of an unbunched (4D) beam with $n = 10^6$ particles sampled from the distribution (67) in the FODO channel shown in Fig. 10. The MMD between the distribution on successive lattice periods $\gamma_k(f_t, f_{t-1})$ decays to near (but remains slightly above) the level of noise (black line). The MMD to the initial distribution $\gamma_k(f_t, f_0)$ is largely unchanged after the first 100 periods.

was used whose width along each dimension matches the initial rms beam size in that dimension. In each case, self-consistent tracking with transverse (2D) space charge using $n = 10^6$ particles was performed using the algorithm described in [63]. The distribution relaxes quickly over the first 100 periods, but fluctuations above the noise level persist on a much longer time scale. This raises the possibility that for $100 < t < 1000$, the distribution may accumulate changes at a rate that is too slow to be resolved by comparing successive periods. However, the fact that $\gamma_k(f_t, f_0)$ is nearly constant over this interval indicates that the size of any such accumulated changes must remain small.

Figure 12 illustrates the time evolution of the parti-

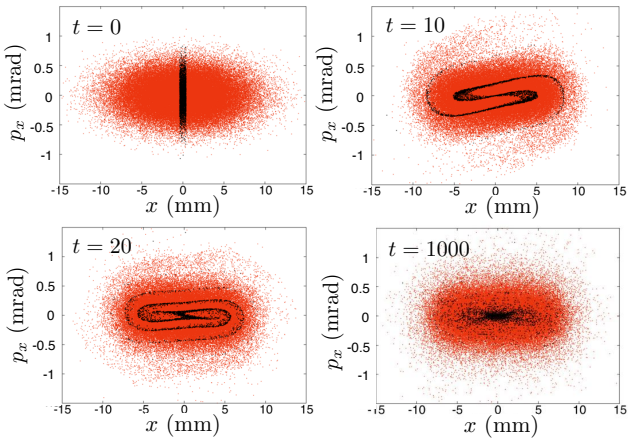

FIG. 12. Projections into the (x, p_x) plane of an initially rms-matched (but non-stationary) distribution (67) evolving in a periodic FODO channel, shown at four distinct lattice periods t. The particles shown are those used to produce Fig. 11. (Red) The complete simulated particle distribution. (Black) A subset of initial conditions, illustrating dynamical filamentation of the phase space.

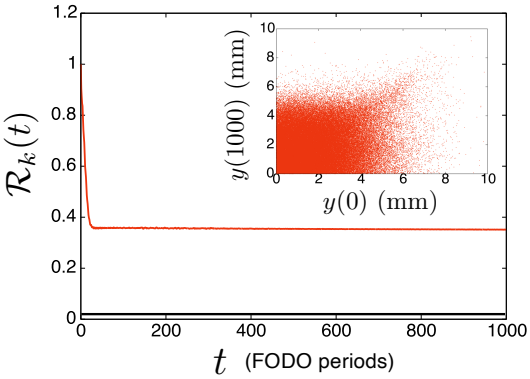

FIG. 13. Dynamics of an unbunched (4D) beam with $n = 10^6$ particles sampled from the distribution (67) in the FODO channel shown in Fig. 10. The correlation $\mathcal{R}_k(t)$ between the distribution after t periods and the initial distribution at $t = 0$ is shown. Correlations appear to persist well-above the level of the noise (black curve). (Inset) Plot of initial y vs. final y after 1K periods, showing that the correlations are not easily visible in low-dimensional projections.

cle distribution in the (x, p_x) plane. A subset of initial conditions is highlighted in black, in order to visualize the dynamics of the phase space points in more detail. The phase space region highlighted in black filaments and folds. Compare this behavior with the dynamics shown in Fig. 9.

Fig. 13 shows the correlation $\mathcal{R}_k(t)$ between the distribution at time t and the distribution at $t = 0$. Note that $\mathcal{R}_k(t)$ converges to a fixed nonzero value within just a few periods, and then remains constant. This indicates that particle coordinates remain correlated with their initial values indefinitely, and that the dynamics is not mixing. This generally suggests the existence of invariants of motion in some regions of the phase space. A 2D plot of $y(0)$ versus $y(1000)$ is also shown. Note that the correlations are not easily visible. In fact, if $\mathcal{R}_k$ is computed using only the initial and final values of y, neglecting all other coordinates, then the corresponding value is 0.08. This shows that $\mathcal{R}_k$ quantifies correlations present in higher dimensions that are not easily visualized by viewing 2D projections, a fact that was also illustrated in Figs. 5-6.

C. Treatment of beam loss

In the presence of beam loss, it may be of interest to study the dynamics of the beam on a bounded subregion E of the phase space (e.g., defined by the vacuum chamber or by the dynamic aperture). For example, one may study the relaxation of the distribution defined by those particles that remain indefinitely within the region E. In this case, the beam distribution function f_t satisfies:

$$\int_E f_t(X)dX = \chi_t, \quad 0 \le \chi_t \le 1, \tag{68}$$

where χ_t denotes the (possibly time-dependent) fraction of beam particles that lie within the region E.

The formalism of the previous section may be modified to treat this case as follows.

- In the calculation of (21-25) and (37-38), sum only over those particles that lie within the desired region E. This is equivalent to modifying the kernel k by setting $k(X, X') = 0$ if $X \notin E$ or $X' \notin E$.
- In the calculation of (21-25) and (37-38), replace the weight coefficients $1/m$ by particle weights w_j with $\sum_{j=1}^{m} w_j = \chi_t$, and similarly for $1/n$.

One may verify that many of the desired mathematical properties of the MMD and HSICor still hold for these modified statistics.

VIII. CONCLUSIONS

Modeling charged particle beams in the presence of high intensity space charge, collective instabilities, or strong nonlinear focusing can result in dynamical processes that are difficult to characterize using typical numerical diagnostics (based on the second-order beam moments or 2-D phase space projections). We have introduced two numerical diagnostics originating in the ML literature with highly desirable mathematical properties that are straightforward to implement in parallel particle-based simulation codes. The first is a measure of statistical distance known as the Maximum Mean Discrepancy

(γ_k), which serves as a measure of similarity between two particle ensembles. The second is a measure of statistical dependence or correlation between random variables known as the Hilbert Schmidt Correlation ($\mathcal{R}_k$).

These quantities are useful for a variety of applications, including: matching a beam into a periodic transport system, numerical benchmarking, detecting possibly nonlinear phase space correlations, characterizing relaxation to a (quasi-)stationary state, and characterizing mixing or decay of correlations within the beam. For these applications, diagnostics based on statistical distance allow one to obtain insight that is not available using conventional diagnostics. As an example, one may quantify fine-grained differences between particle distributions, which may occur in degrees of freedom that are not easily visualized, and one may study the evolution of 4-D or 6-D nonlinear correlations among phase space variables within the beam, which are often not visible in its 2-D phase space projections.

It is important to note that the quantitative results obtained will depend on the choice of the kernel k. On one hand, this kernel-dependence may be viewed as a disadvantage of the diagnostics described here. On the other hand, one may view the choice of kernel as a natural way to parameterize a large family of possible diagnostics, all of which correctly capture the same underlying physical processes. (This is a consequence of the mathematical properties described in Sections IV-V.) In numerical experiments the observed dynamical evolution of γ_k or $\mathcal{R}_k$ was largely independent of the choice of kernel, although this remains an active area of investigation. An alternative and parameter-free statistical distance with similar mathematical properties is the Wasserstein distance W_p (Section III B). In the future, the authors plan to investigate the feasibility of using efficient approximations to W_p [64] for beam dynamics applications.

The diagnostics described here are well-suited to applications involving large simulation ensembles. For example, quantities involving γ_k or $\mathcal{R}_k$ may be used as objectives for accelerator design optimization or for training machine learning models that require detailed information about the beam distribution function. This raises the possibility of tailoring the final beam phase space density by using large-scale automated machine tuning.

Finally, although we have focused on the case of charged particle beams, it is clear that these tools can be applied without change to kinetic simulations of other many-body systems such as plasmas or gravitational systems, which rely on the tracking of large particle ensembles.

IX. ACKNOWLEDGMENTS

This work was supported by the Director, Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and made use of computer resources at the National Energy Research Scientific Computing Center. The authors acknowledge support from the U.S. DOE Early Career Research Program under the Office of High Energy Physics.

APPENDIX A: COMMONLY-USED KERNELS

This Appendix lists several of the kernels k most commonly used for discrimination testing and independence testing in ML. These kernels are all translation-invariant and normalized as in (13), so that $k(X, X) = 1$ for all X. All the kernels listed here have the property that the quantity γ_k in (12) satisfies the metric conditions (i-iv) and captures weak convergence, as described in Section III.

Gaussian kernel ($\sigma > 0$) of dimension d:

$$k_{\text{Gaussian}}(X, X') = e^{-|X-X'|^2/2\sigma^2}, \tag{69a}$$

$$\Lambda_{\text{Gaussian}}(\omega) = \frac{e^{-\sigma^2|\omega|^2/2}}{(2\pi\sigma^2)^{d/2}}. \tag{69b}$$

Laplace kernel ($\sigma > 0$) of dimension d:

$$k_{\text{Laplace}}(X, X') = e^{-|X-X'|/\sigma}, \tag{70a}$$

$$\Lambda_{\text{Laplace}}(\omega) = \frac{\sigma^d}{\pi^{(d+1)/2}} \frac{\Gamma\left(\frac{d+1}{2}\right)}{(1+\sigma^2\omega^2)^{(d+1)/2}}, \tag{70b}$$

where Γ denotes the gamma function.

Matérn kernel ($\nu > 0$, $\sigma > 0$) of dimension d:

$$k_{\text{Matern}}(X, X') = \frac{2^{1-\nu}}{\Gamma(\nu)} \zeta^\nu K_\nu(\zeta), \tag{71a}$$

$$\Lambda_{\text{Matern}}(\omega) = \frac{2^s \sigma^d \nu^\nu \Gamma(s)}{(2\pi)^{d/2}\Gamma(\nu)} \left(2\nu + \sigma^2 |\omega|^2\right)^{-s}, \tag{71b}$$

where K_ν is the modified Bessel function of order ν, and we abbreviate:

$$\zeta = \frac{\sqrt{2\nu}}{\sigma}|X - X'|, \qquad s = \nu + \frac{d}{2}. \tag{72}$$

Poisson kernel ($0 < \sigma < 1$) of dimension 1:

$$k_{\text{Poisson}}(X, X') = \frac{(1-\sigma)^2}{1 - 2\sigma\cos(X - X') + \sigma^2}, \tag{73a}$$

$$\Lambda_{\text{Poisson}}(\omega) = \sum_{n=-\infty}^{\infty} \left(\frac{1-\sigma}{1+\sigma}\right) \sigma^{|n|} \delta(\omega - n). \tag{73b}$$

Another class of kernels k_{Wendland} (of dimension d) is constructed by using a polynomial with compact support in the separation distance $r = |X - X'|/\sigma$, where $\sigma > 0$. See [65, 66] for a detailed description of these. An example for $d = 1$ is given by:

$$k_{\text{Wend}}(X, X') = (1-r)^3(1+3r), \qquad r < 1, \tag{74a}$$

$$\Lambda_{\text{Wend}}(\omega) = \frac{24\sigma}{\pi} \left\{ \frac{2 + \cos\omega\sigma}{(\omega\sigma)^4} - \frac{3\sin\omega\sigma}{(\omega\sigma)^5} \right\}. \tag{74b}$$

When working with a Gaussian kernel, an orthonormal basis for the RKHS is given by $\{e_n\}_{n=0}^{\infty}$ where [67]:

$$e_n(X) = \sqrt{\frac{1}{\sigma^{2n} n!}} e^{-X^2/2\sigma^2} X^n. \tag{75}$$

When working with the Poisson kernel, a (complex) orthonormal basis for the RKHS is given by $\{e_n\}_{n=-\infty}^{\infty}$ with:

$$e_n(X) = \sqrt{\frac{1-\sigma}{1+\sigma}} \sigma^{|n|/2} e^{inX}. \tag{76}$$

A corresponding real basis is easily constructed.

We can construct a kernel k of higher dimension using kernels $k^{(j)}$ $(j = 1, \ldots, d)$ of lower dimension. For example, if we write $X = (X_1, \ldots, X_d)$ and $X' = (X'_1, \ldots, X'_d)$, then

$$k(X, X') = \prod_{j=1}^{d} k^{(j)}(X_j, X'_j). \tag{77}$$

If the kernels $k^{(j)}$ are translation-invariant with spectral densities $\Lambda^{(j)}$, then setting $\omega = (\omega_1, \ldots, \omega_d)$ we have:

$$\Lambda(\omega) = \prod_{j=1}^{d} \Lambda^{(j)}(\omega_j). \tag{78}$$

If the RKHS associated with the kernel $k^{(j)}$ has basis $e_n^{(j)}$, then the RKHS associated with k has basis

$$e_n(X) = \prod_{j=1}^{d} e_{n_j}^{(j)}(X_j), \tag{79}$$

where $n = (n_1, \ldots, n_d)$ ranges over all possible indices.

APPENDIX B: DEFINITION OF MIXING

Let $\mathcal{M}$ be a map, and let f denote a probability density that is invariant under the map, in the sense that:

$$f(\mathcal{M}^{-1}(X)) = f(X). \tag{80}$$

Then the map $\mathcal{M}$ is mixing with respect to the invariant density f if for any two sets A and B [68]:

$$\lim_{t\to\infty} P(\mathcal{M}^{-t}(A) \cap B) = P(A)P(B), \tag{81}$$

where

$$P(A) = \int_A f(X) dX \tag{82}$$

denotes the probability that a point lies in A. Informally, for any sets A and B, the sequence of sets $\mathcal{M}^{-t}(A)$ becomes asymptotically independent of B as $t \to \infty$ [59].

There are many equivalent formulations of the condition (81). For our purposes, it is enough to know that a map $\mathcal{M}$ is mixing with respect to a density f if and only if $\mathcal{R}_k(t) \to 0$ as $t \to \infty$, where $\mathcal{R}_k(t)$ denotes the correlation between the particle coordinates sampled from f at $t = 0$ and the corresponding particle coordinates at later time t. (See Section V.) This holds for all kernels k satisfying the desirable properties described in Section III C.

APPENDIX C: ERROR BOUNDS

We provide an upper bound for the error associated with the estimates (21) and (25), in order to clarify the dependence on the parameters n, m, (number of particles) and L (number of frequency samples).

Define a quantity $\Delta\gamma_{\text{rms}}$ by:

$$\Delta\gamma_{\text{rms}} = \mathrm{E}[\{\gamma_k^{\text{sample}}(f, g) - \gamma_k(f, g)\}^2]^{1/2}, \tag{83}$$

where γ_k^{sample} denotes the estimated value computed from samples using (25), γ_k denotes the exact value given by the integral (12), and E denotes the expected value taken by averaging over independently sampled realizations of f, g, and Λ.

Using the the triangle inequality property iii) of γ_k, the Cauchy-Schwarz inequality, and basic properties of expectation and variance, one obtains an upper bound for $\Delta\gamma_{\text{rms}}$ given by:

$$\begin{aligned}\Delta\gamma_{\text{rms}} \leq &\left[\frac{1}{n}\left(1 - ||f||_k^2\right) + \frac{1}{m}\left(1 - ||g||_k^2\right)\right]^{1/2} \\ &+ \min\left\{\sqrt{2}\gamma_k(f, g), \frac{1}{\sqrt{L}} d_{TV}(f, g)\right\}.\end{aligned} \tag{84}$$

The meaning of $||\cdot||_k$ was defined in (27), while d_{TV} denotes the total variation distance between f and g, defined here by:

$$d_{TV}(f, g) = \int |f(X) - g(X)| \, dX \leq 2. \tag{85}$$

We make the following observations:

1) The first line of (84) denotes the contribution to the error due to the finite number of particles, which scales as $O(n^{-1/2} + m^{-1/2})$.

2) In the special case that $||f||_k = ||g||_k = 1$, the first line of (84) vanishes. This occurs, for example, if f and g are each concentrated at a single point (*i.e.*, they are Dirac-delta functions).

3) The second line of (84) denotes the contribution to the error due to the finite number of frequency samples, which scales as $O(L^{-1/2})$ when $L \to \infty$. The coefficient $d_{TV}(f, g)$ is small when f and g are similar.

4) When $\gamma_k(f, g) \to 0$, the error contribution due to frequency sampling vanishes, and the inequality can be replaced by equality, yielding (26).

5) In the limit $L \to \infty$, (84) yields an upper bound for the rms error associated with the estimate (21).

A numerical example illustrating the behavior of $\Delta\gamma_{\rm rms}$ is shown in Fig. 1.

[1] H. Kandrup *el al*, "Chaotic collisionless evolution in galaxies and charged-particle beams," Ann. N. Y. Acad. Sci. **1045**, 12 (2005).

[2] C. Bohn, "Chaotic dynamics in charged-particle beams: possible analogs of galactic evolution," Ann. N. Y. Acad. Sci. **1045**, 34 (2005).

[3] D. Stratakis *et al*, "Experimental and numerical study of phase mixing of an intense beam," Phys. Rev. ST Accel. Beams **12**, 064201 (2009).

[4] Y. Levin *et al*, "Nonequilibrium statistical mechanics of systems with long-range interactions," Phys. Rep. **535**, 1-60 (2014).

[5] B. B. Kadomtsev and O. P. Pogutse, "Collisionless relaxation in systems with Coulomb interactions," Phys. Rev. Lett. **25**, 1155 (1970).

[6] Y. Levin, R. Pakter, and T. Teles, "Collisionless Relaxation in Non-Neutral Plasmas," Phys. Rev. Lett. **100**, 040604 (2008).

[7] A. Gretton *et al*, "A Kernel Two-Sample Test", Journal of Machine Learning Research **13**, 723-773 (2012).

[8] T. Hofmann, B. Scholkopf, and A. J. Smola, "Kernel Methods in Machine Learning," The Annals of Statistics **36**, 1171-1220 (2008).

[9] V. Paulsen and M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge University Press, 2016.

[10] A. Gretton *et al*, "Measuring Statistical Dependence with Hilbert-Schmidt Norms", Technical report no. 140, Max Planck Institute for Biological Cybernetics (2005), `http://www.gatsby.ucl.ac.uk/~gretton/papers/GreHerSmoBouSch05a.pdf`.

[11] A. Campa, T. Dauxois, and S. Ruffo, "Statistical mechanics and dynamics of solvable models with long-range interactions," Physics Reports **480**, 57-159 (2009).

[12] Y. Levin *et al*, "Nonequilibrium statistical mechanics of systems with long-range interactions," Physics Reports **535**, 1-60 (2014).

[13] A. Campa *et al*, *Physics of Long-Range Interacting Systems*, Oxford Academic, Oxford, 2014.

[14] R. L. Liboff, *Kinetic Theory: Classical, Quantum, and Relativistic Descriptions*, 3rd ed., Springer-Verlag, New York (2003).

[15] Pierre-Henri Chavanis, "Kinetic theory of collisionless relaxation for systems with long-range interactions," Physica A **606**, 128089 (2022).

[16] C. Mouhot and C. Villani, "On Landau Damping," Acta Math. **207**, 29 (2011).

[17] C. Villani, "Particle systems and nonlinear Landau damping," Phys. Plasmas **21**, 030901 (2014).

[18] S. Jaffard *et al*, "Multifractal formalisms for multivariate analysis," Proc. R. Soc. A **475**, 20190150 (2019).

[19] C. W. van Atta and W. Y. Chen, "Structure functions of turbulence in the atmospheric boundary layer over the ocean," J. Fluid Mech. **44**, 145-159 (1970).

[20] O. N. Boratav and R. B. Pelz, "Structures and structure functions in the intertial range of turbulence," Physics of Fluids **9**, 1400 (1997).

[21] R. T. Cerbus and W. I. Goldburg, "Information content of turbulence," Phys. Rev. E **88**, 053012 (2013).

[22] L. Chevillard, B. Castaing, and E. Lévêque, "On the rapid increase of intermittency in the near-dissipation range of fully developed turbulence," Eur. Phys. J. B **45**, 561-567 (2005).

[23] M. Reiser, *Theory and Design of Charged Particle Beams*, 2nd ed, Wiley-VCH, Weinheim, 2008.

[24] R. C. Davidson and H. Qin, *Physics of Intense Charged Particle Beams in High Energy Accelerators*, World Scientific, River Edge, NJ, 2001.

[25] By a "probability density" we mean, more generally, a probability measure defined on the σ-algebra of Borel subsets of the phase space $\mathbb{R}^d$. Such a measure may not possess a density in the strict sense of a function in $L^1(\mathbb{R}^d)$. For example, our "probability density" here is allowed to contain Dirac delta functions.

[26] R. M. Dudley, *Real Analysis and Probability*, Cambridge Studies in Advanced Mathematics 74, Cambridge University Press, 2002.

[27] P. Billingsley, *Convergence of Probability Measures*, 2nd ed, Wiley, NY, 1999.

[28] M. Markatou, D. Karlis, and Y. Ding, "Distance-Based Statistical Inference," Annu. Rev. Stat. Appl. **8**, 301-27 (2021).

[29] S. Kullback and R. Leibler, "On Information and Sufficiency," Annals of Mathematical Statistics **22**, 79-86 (1951).

[30] S. Kullback, *Information Theory and Statistics*, Dover Publications, Inc., Mineola, NY (1968).

[31] C. Granero-Belinchon, S. Roux, and N. Garnier, "Kullback-Leibler divergence measure of intermittency: Application to turbulence," Phys. Rev. E **97**, 013107 (2018).

[32] Q. Wang *et al*, "Divergence Estimation for Multidimensional Densities via k-Nearest-Neighbor Distances", IEEE Transactions on Information Theory **55**, 2392 (2009).

[33] K. Ahuja, "Estimating Kullback-Leibler Divergence Using Kernel Machines," arXiv:1905.00586v2 (2019).

[34] Strictly speaking, the infimum of this quantity must be taken over all "joint" probability measures defined on the Borel subsets of $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals have densities f and g, respectively. See comment [25].

[35] L. V. Kantorovich, "On the Translocation of Masses", Dokl. Akad. Nauk USSR **37**, pp. 227-229 (1942). [English translation: Journal of Mathematical Sciences **133**, 1381-1382 (2006)].

[36] L. N. Wasserstein, "Markov Processes Over Denumerable Products of Spaces Describing Large Systems of Automata," Probl. Inform. Transmission **5**, pp. 47-52 (1969).

[37] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2003.

[38] P. Komiske, E. Metodiev, and J. Thaler, "Metric Space

of Collider Events," Phys. Rev. Lett. **123**, 041801, 2019.
[39] M. Erdmann *el*, "Generating and Refining Particle Detector Simulations Using the Wasserstein Distance in Adversarial Networks," Computing and Software for Big Science 2:4 (2018).
[40] T. Cai, J. Cheng, and N. Craig, "Linearized Optimal Transport for Collider Events," Phys. Rev. D **102**, 116019 (2020).
[41] B. Sriperumbudur *et al*, "On the Empirical Estimation of Integral Probability Metrics", Electronic Journal of Statistics **6**, 1550-1599 (2012).
[42] B. Sriperumbudur *et al*, "Non-parametric Estimation of Integral Probability Metrics", in Proc. 2010 IEEE International Symposium on Information Theory, 13-18 June 2010.
[43] B. Sriperumbudur *et al*, "Hilbert Space Embeddings and Metrics on Probability Measures", Journal of Machine Learning Research **11**, 1517-1561 (2010).
[44] B. Sriperumbudur, K. Fukumizu, and G. Lanckriet, "Universality, Characteristic Kernels and RKHS Embeddings of Measures," Journal of Machine Learning Research **12**, 2389-2410 (2011).
[45] B. Sriperumbudur, "On the Optimal Estimation of Probability Measures in the Weak and Strong Topologies," Bernoulli **22**, 1839-1893 (2016).
[46] C. J. Simon-Gabriel and B. Schölkopf, "Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions," Journal of Machine Learning Research **19**, 1-29 (2018).
[47] C. J. Simon-Gabriel *et al*, "Metrizing Weak Convergence with Maximum Mean Discrepancies" (2021), `https://arxiv.org/abs/2006.09268`.
[48] D. Sejdinovic *et al*, "Equivalence of Distance-Based and RKHS-Based Statistics in Hypothesis Testing", The Annals of Statistics **41**, 2263-2291 (2013).
[49] G. J. Székely and M. L. Rizzo, "Energy statistics: A class of statistics based on distances," Journal of Statistical Planning and Inference **143**, 1249-1272 (2013).
[50] J. Zhao and D. Meng, "FastMMD: Ensemble of circular discrepancy for efficient two-sample test", Neural Computation **27**, 1345-1372 (2015).
[51] A. Rahimi and B. Recht, "Random Features for Large-Scale Kernel Machines," Adv. Neural Inf. Process. Syst. **20** (2007).
[52] A. Gretton *et al*, "Measuring Statistical Dependence with Hilbert-Schmidt Norms", Lecture Notes in Computer Science, pp. 63-77 (2005).
[53] A. Gretton *et al*, "A Kernel Statistical Test of Independence", in Proc. of the 20th International Conference on Neural Information Processing Systems, p. 585-592 (2007).
[54] G. Székely, G. Rizzo, and M. Bakirov, "Measuring and testing dependence by correlation of distances", Ann. Stat. **35**, 2769-2794 (2007).
[55] R. Lyons, "Distance Covariance in Metric Spaces", The Annals of Probability **41**, 3284-3305 (2013).
[56] Q. Zhang *et al*, "Large-scale kernel methods for independence testing", Stat. Comput. **28**, 113-130 (2018).
[57] R. E. Meller *et al*, "Decoherence of Kicked Beams," Technical Note SSC-N-360 (1987), `https://lss.fnal.gov/archive/other/ssc/ssc-n-360.pdf`
[58] C. Mitchell, "Weak Convergence to Equilibrium of Statistical Ensembles in Integrable Hamiltonian Systems", Journal of Mathematical Physics **60**, 052702 (2019).
[59] P. Walters, *An Introduction to Ergodic Theory*, Springer-Verlag, New York, 1982.
[60] V. I. Arnold and A. Avez, *Ergodic Problems of Classical Mechanics* (Benjamin: New York, 1968).
[61] Most authors define the map (58) with $\mod 2\pi$ replaced by $\mod 1$. This amounts to a choice of coordinates on the torus $\mathbb{T}^2$. The resulting map is equivalent to (58), after q and p are each scaled by 2π.
[62] K. T. Alligood, Tim D. Sauer, and James A. Yorke, *Chaos: an Introduction to Dynamical Systems*, Springer-Verlag New York, NY, 1997.
[63] J. Qiang, "Symplectic multiparticle tracking model for self-consistent space-charge simulation," Phys. Rev. Accel. Beams **20**, 014203 (2017).
[64] S. Nietert, Z. Goldfeld, K. Kato, "Smooth p-Wasserstein distance: structure, empirical approximation, and statistical applications," Proceedings of the 38th International Conference on Machine Learning, PMRL 139, 8172 (2021).
[65] H. Wendland, "Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree," Adv. Comput. Math. **4**, 389-396 (1995).
[66] A. Chernih and S. Hubbert, "Closed form representations and properties of the generalised Wendland functions," J. Approx. Theory **177**, 17-33 (2014).
[67] H. Minh, "Some Properties of Gaussian Reproducing Kernel Hilbert Spaces and Their Implications for Function Approximation and Learning Theory", Constructive Approximation **32**, 307-338 (2010).
[68] This is the definition of *strong mixing*. It is assumed that A and B are "nice" (Borel) subsets of the phase space $\mathbb{R}^d$, and the map $\mathcal{M}$ is Borel-measurable. This holds, for example, if $\mathcal{M}$ is continuous.