Sampling from Arbitrary Functions via PSD Models

Ulysse Marteau-Ferey, Francis Bach, Alessandro Rudi

Sampling from Arbitrary Functions via PSD Models

Ulysse Marteau-Ferey, Francis Bach, Alessandro Rudi

INRIA - Département d'Informatique de l'École Normale Supérieure
PSL Research University
Paris, France

February 23, 2022

Abstract

In many areas of applied statistics and machine learning, generating an arbitrary number of independent and identically distributed (i.i.d.) samples from a given distribution is a key task. When the distribution is known only through evaluations of the density, current methods either scale badly with the dimension or require involved implementations. Instead, we take a two-step approach by first modeling the probability distribution and then sampling from that model. We use the recently introduced class of positive semi-definite (PSD) models, which have been shown to be efficient for approximating probability densities. We show that these models can approximate a large class of densities concisely using few evaluations, and present a simple algorithm to effectively sample from these models. We also present preliminary empirical results to illustrate our assertions.

1 Introduction

In many fields such as biochemistry, statistical mechanics and machine learning, effectively sampling arbitrary numbers of independent and identically distributed (i.i.d.) samples from probability distributions is a key task [4, 6, 5].

Basic sampling methods include rejection sampling and gridding, and rely on simple properties of the density. However, they are suitable only in small dimensions, except for very structured cases. Moreover, they are hard to adapt to probabilities which are known up to their renormalization constant, which is often the case when dealing with exponential models that are common in applications [12].

More involved methods have been developed to address these dimensionality and renormalization issues, in the class of so-called Markov chain Monte Carlo (MCMC) methods. However, they are complex to set up: in particular, independence between samples is not directly guaranteed, convergence can be slow and hard to measure non-asymptotically [5, 12].

In this work, we address the problem in a different way, by incorporating a modeling step. Instead of sampling directly from the target density, we first model this density using a positive semi-definite (PSD) model [9, 15], and then sample from this PSD model.

PSD models have been introduced by Marteau-Ferey et al. [9] and their relevance for modeling probability distributions has been further established by Rudi and Ciliberto [15], showing that i) they are stable under key operations for probabilistic inference, such as marginalization, integration (also called "sum-rule"), and product, which can be done efficiently in practice, and ii) they concisely approximate a large class of probability distributions. We present these models in Sec. 2. Building

on this work, we show that these models are also relevant in the context of sampling, making the following main contributions.

(1) In Sec. 3, we derive an algorithm that is easy to implement and which can generate an arbitrary number of i.i.d. samples from a given PSD model, with any given precision. This answers one of the open questions outlined by Rudi and Ciliberto [15] and shows that one can indeed efficiently sample from a PSD model.

(2) In Sec. 4 we show that we can sample an arbitrary number of i.i.d. samples from a target probability distribution that is regular enough, with any given precision. The algorithm consists in (a) approximating the un-normalized density p via a PSD model, using evaluations of p, and (b) extracting i.i.d. samples from the PSD model. We show that for sufficiently regular densities the resulting PSD model is concise and avoids the curse of dimensionality: to achieve error ε, the PSD model requires a number of parameters and a number of evaluations of p that are in the order $\varepsilon^{-2-d/\beta}$, where d is the dimension of the space and β is the order of differentiability of the density. For regular probabilities, i.e., when $\beta \geqslant d$, the rate does not depend exponentially on d and is bounded by $O(\varepsilon^{-3})$ (the constant term instead may depend exponentially on d).

In Sec. 5, we also present numerical simulations which demonstrate the quality of both our sampling technique and approximation results.

2 Backround on Positive Semi-Definite (PSD) models

Denote by $\mathbb{R}^d_{++}$ the vectors of $\mathbb{R}^d$ with positive components and $\mathbb{S}^m_+$ the set of positive semi-definite m by m matrices. Following Marteau-Ferey et al. [9], Rudi and Ciliberto [15], a Gaussian PSD model is parametrized by a triplet $(A, X, \eta) \in \mathbb{S}^m_+ \times \mathbb{R}^{m\times d} \times \mathbb{R}^d_{++}$, and is defined for any $x \in \mathbb{R}^d$ as

$$f(x;\ A, X, \eta) = \sum_{i,j=1}^{m} A_{ij} k_\eta(x, x_i) k_\eta(x, x_j), \tag{1}$$

where, with $\mathrm{diag}(\eta)$ being the diagonal matrix with diagonal η, $k_\eta(x, x') = e^{-(x-x')^\top \mathrm{diag}(\eta)(x-x')}$ is the Gaussian kernel of parameter η , $X \in \mathbb{R}^{n\times d}$ is the matrix whose rows corresponds to the centers $x_1, \ldots, x_n$ of the Gaussian PSD model, and A is a matrix of coefficients which is positive semi-definite, to guarantee the non-negativity of f.

Note that when $A = aa^\top,\ a \in \mathbb{R}^m$, is a rank-1 operator, a Gaussian PSD model is simply the square of a linear model $f(x;\ A, X, \eta) = g(x;\ a, X, \eta)^2$ of the form,

$$g(x;\ a, X, \eta) = \sum_{i=1}^{m} a_i k_\eta(x, x_i), \tag{2}$$

for any $x \in \mathbb{R}^d$. This particular case of PSD model will appear when approximating an arbitrary probability density p in Sec. 4.2.

2.1 Main properties of PSD models

As explained in the introduction, PSD models show properties that make them particularly well suited to model non-negative functions and probability distributions. Such properties are analyzed by Marteau-Ferey et al. [9] and Rudi and Ciliberto [15], here we recall the ones that are important for our purpose.

Non-negativity. Since A is positive semidefinite, then the PSD model $f(x;\ A, X, \eta)$ satisfies $f(x;\ A, X, \eta) \geqslant 0$ for all $x \in \mathbb{R}^d$.

Preservation of convex functionals. Using the PSD model to represent non-negative functions in a problem of the form $\min_{f \geqslant 0} L(f)$, where L is a convex functional, leads to a convex problem $\min_{A \in \mathbb{S}_+(\mathbb{R}^m)} L(f(\cdot;\ A, X, \eta))$. Indeed, the constraint $A \in \mathbb{S}_+(\mathbb{R}^m)$ is convex, the PSD model $f(\cdot;\ A, X, \eta)$ is linear in the parameter matrix A and a composition of a convex function L with a linear function is convex. This allows, e.g., to perform empirical risk minimization for the square and logarithmic losses.

Conciseness of the representation. under mild conditions, recalled in Assumption 1, a PSD model can approximate a probability density that is β-times differentiable with error ε, using a number of centers $m = O(\varepsilon^{-d/\beta})$ (which is minimax optimal). Rudi and Ciliberto [15] provide also an algorithm to learn the PSD model given i.i.d. samples from the probability. However, we cannot use this result in our context since we do not assume to have samples from our density.

Integration over hyper-rectangles in closed form. As integration of PSD models will play a key role in the algorithm developed for sampling in Sec. 3, both for theoretical an computational reasons, we develop this integration aspect in greater detail.

A hyper-rectangle $Q \subset \mathbb{R}^d$ can be parametrized with its corners $a, b \in \mathbb{R}^d,\ a \leqslant b$, by writing $Q = \prod_{k=1}^d [a_k, b_k[$; a corresponds to the "bottom left" corner and b to the "top right" one.

For $X \in \mathbb{R}^{m \times d}$ and $\eta \in \mathbb{R}^d_{++}$, we denote with $K_{X,\eta} \in \mathbb{R}^{m \times m}$ the *kernel matrix* such that $[K_{X,\eta}]_{ij} = k_\eta(x_i, x_j)$. The integral of a PSD model in Eq. (1) over a hyper-rectangle can be expressed with simple matrices, leveraging the fact that for any pair (x_i, x_j), it holds $k_\eta(x, x_i)k_\eta(x, x_j) = k_{\eta/2}(x_i, x_j)k_{2\eta}(x, (x_i + x_j)/2)$. Then we have

$$\begin{aligned} I(Q; A, X, \eta) &:= \int_Q f(x; A, X, \eta)\ dx \\ &= \sum_{i,j=1}^m A_{ij} k_{\frac{\eta}{2}}(x_i, x_j) \int_Q k_{2\eta}(x, \tfrac{x_i + x_j}{2})\ dx \\ &= \sum_{i,j=1}^m A_{ij} [K_{X,\eta/2}]_{ij} [G_{X,2\eta,Q}]_{ij}, \end{aligned} \tag{3}$$

where $[G_{X,\eta,Q}]_{ij} = \int_{Q_{ij}} k_\eta(x, 0)\ dx$, and $Q_{ij} = Q - (x_i + x_j)/2$. These integrals can be computed by $2d$ calls to the erf function, as, for any $i, j \in \{1, ..., m\}$:

$$[G_{X,\eta,Q}]_{ij} = c_\eta \prod_{k=1}^d [\mathrm{erf}(\sqrt{\eta_k}\mathcal{B}_{ijk}) - \mathrm{erf}(\sqrt{\eta_k}\mathcal{A}_{ijk})], \tag{4}$$

where $c_\eta = (\pi/4)^{d/2} \det \mathrm{diag}(\eta)^{-1/2}$, $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{d \times m \times m}$, $\mathcal{A}$ is the tensor of bottom left corners and $\mathcal{B}$ is the tensor of top right corners, defined formally from the means tensor $\overline{X}_{ijk} = \frac{1}{2}(X_{ik} + X_{jk})$ as

$$\mathcal{A}_{ijk} = a_k - \overline{X}_{ijk}, \quad \mathcal{B}_{ijk} = b_k - \overline{X}_{ijk}. \tag{5}$$

This shows that, for any hyper-rectangle Q, we can compute $G_{X,\eta,Q}$ with exactly $2dm^2$ calls to the erf function and dm^2 arithmetic operations (so there is no dependence on the dimension of the hyper-rectangle).

3 A sampling algorithm for PSD models

In this section, we fix a Gaussian PSD model on $\mathbb{R}^d$ parametrized by $(A, X, \eta) \in \mathbb{S}_+^m \times \mathbb{R}^{m\times d} \times \mathbb{R}_{++}^d$ for a given $m \in \mathbb{N}$. To simplify notations, we will omit the parameters of the PSD model using $f(x)$ as a shorthand for $f(x; A, X, \eta)$ and $I(Q)$ as a shorthand of $I(Q) = I(Q; A, X, \eta)$.

Given a bounded hyper-rectangle Q (see Sec. 3.1), denote by p_Q the function

$$p_Q(x) = f(x)\mathbf{1}_Q(x)/I(Q), \tag{6}$$

where $\mathbf{1}_Q(x) = 1$ when $x \in Q$ and 0 otherwise. In Sec. 3.2, we explain that even in the case of an infinite hyper-rectangle (e.g., $Q = \mathbb{R}^d$), we can easily find a finite hyper-rectangle $\tilde{Q}$ on which the whole mass of f is essentially concentrated, and thus approximately sample in this case as well. We end this section with a discussion on the main elements needed to sample, and which could allow to generalize this approach to PSD models with different kernels.

3.1 A sampling algorithm on a finite hyper-rectangle

Given the function f, the algorithm will take three inputs (Q, N, ρ): the hyper-rectangle Q (with sides parallel to the axes) from which we would like to sample, the number of i.i.d. samples N which we would like to obtain, and a parameter ρ which defines the quality of the approximation of p_Q from which the algorithm generates samples. The effect of ρ on the precision of the algorithm is formally established in Theorem 2.

We start with the case $N = 1$. Starting from Q, we cut Q in half in its longest direction forming two sub-rectangles Q_1, Q_2. If X_Q were a random variable following the law of p_Q, then $X_Q \in Q_i$ with probability $p_i = I(Q_i)/I(Q)$, and $X_Q|\{X_Q \in Q_i\}$ follows the law of p_{Q_i}. Therefore, when looking for a sample from p_Q, we randomly choose with probability p_i one of the two smaller sub-rectangles Q_i in which to look for the sample and then call the algorithm recursively to get a sample from p_{Q_i}. Of course, we need a stopping criterion: when the maximal side of Q has length smaller than ρ then we stop and we return a point sampled uniformly at random in Q. The complete algorithm is presented in algorithm 1 and is explained below.

Details for algorithm 1. In algorithm 1, we define the recursive function SampleRec which will generate samples recursively. The main algorithm Sample in algorithm 1 simply calls the function SampleRec and randomly reshuffles the samples in order to guarantee independence (see RandomPerm algorithm 1). In algorithm 1, the function MaxLen applied to Q returns the maximum of the lengths of the sides of Q; the condition can therefore be translated as "if all sides of Q are smaller than ρ". If it is the case, in algorithm 1, we return N i.i.d. samples from the uniform distribution on Q using SampleUniform. If it is not, in line algorithm 1 we cut the hyper-rectangle Q in half along its largest side with minimal index (i.e., along side $k = \min \operatorname{argmax}(b_i - a_i)$), yielding two sub hyper-rectangles Q_1, Q_2. This is the purpose of the function SplitLargestSide. In algorithm 1, we compute the probability q that a given sample from p_Q belongs to Q_1 using the fact that we can integrate the PSD model exactly. Since we have to generate N samples, we will select k of them from Q_1 and $N - k$ from Q_2 where k is a sample from a binomial law of paramter q: this is the purpose of the function SampleBinomial and algorithm 1. We then call the algorithm recursively to generate the k samples from Q_1 using p_{Q_1} and the $N - k$ samples from Q_2 from p_{Q_2} (algorithm 1).

Guarantees of the algorithm. Given (Q, N, ρ), algorithm 1 does not sample N i.i.d. samples from the exact distribution p_Q but rather from an approximation $p_{Q,\rho}$ of p_Q, controlled by the

Algorithm 1 Approximately sampling from p_Q

function SAMPLEREC(Q, N, ρ)
  if $N = 0$ **then**
    return EMPTYLIST
  else if MAXLEN(Q) $\leqslant \rho$ **then**
    return SAMPLEUNIFORM(Q, N)
  else
    $Q_1, Q_2 =$ SPLITLARGESTSIDE(Q)
    $q = I(Q_1)/I(Q)$
    $k =$ SAMPLEBINOMIAL(N, q)
    $L_1 =$ SAMPLEREC(Q_1, k, ρ)
    $L_2 =$ SAMPLEREC($Q_2, N - k, \rho$)
    return CONCATENATE(L_1, L_2)
  end if
end function

function SAMPLE(Q, N, ρ)
  $L =$ SAMPLEREC(Q, n, ρ)
  return RANDOMPERM(L)
end function

parameter ρ. More formally, let $\mathcal{D}_{Q,\rho}$ be the set of dyadic sub-rectangles of Q with largest possible size smaller than ρ (see Appendix D for a formal definition). Our algorithm will effectively sample from a piece-wise constant approximation of p on the elements of $\mathcal{D}_{Q,\rho}$:

$$p_{Q,\rho} = \frac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \frac{I(Q_\rho)}{|Q_\rho|} \mathbf{1}_{Q_\rho}, \tag{7}$$

where $\mathbf{1}_{Q_\rho}$ is the indicator function of Q_ρ. The guarantees of the algorithm are established in the following theorem, proved formally in Appendix D.2.

Theorem 1. *Given (Q, N, ρ) where Q is a bounded hyper-rectangle of $\mathbb{R}^d$, $\rho > 0$ and $N \in \mathbb{N}$, the function* SAMPLE *in algorithm 1 returns N i.i.d. samples from the distribution $p_{Q,\rho}$ defined in Eq. (7). Moreover, the number of integral computations of the form $I(\widetilde{Q})$ performed during the algorithm is bounded by $N \log_2(|Q|) + Nd \log_2 \frac{2}{\rho} + 1$, and the number of* erf *computations is $O(N\ m^2\ d\ (\log_2(2|Q|) + d \log_2(2/\rho)))$, where m is the dimension of the PSD model.*

Note that the theorem gives us that the complexity is essentially $O(Nm^2d^2 \log(1/\rho))$. This quadratic dependence in the dimension d is verified in practice and the slicing procedure does not yield any time or computational difficulties. Note however that in our two step procedure detailed in the next section, the number m will a priori depend on the dimension, but this is confined to the learning phase; once the m centers are set, the complexity is quadratic. Moreover, note that we verify the claim that computing integrals is the computational bottleneck in practice in Appendix D.4.

Approximation error of the algorithm. Since by Theorem 1, the algorithm does not generate samples exactly from p_Q but rather from the piecewise constant approximation $p_{Q,\rho}$ defined in Eq. (7), it is necessary to quantify the distance between p_Q and its approximation $p_{Q,\rho}$. We do so in Theorem 2 for three different distances.

The weakest distance will be the Wasserstein-1 distance (also called earth mover's distance) [17]. It quantifies the discrepancies in the allocation of mass between two distributions, and is defined

as

$$\mathbb{W}_1(p_1, p_2) = \sup_{\mathrm{Lip}(f)\leqslant 1} \left| \int_{\mathcal{X}} f(x)(p_1(x) - p_2(x))dx \right|, \tag{8}$$

where $\mathrm{Lip}(f)$ is the Lipschitz constant of f for the Euclidean norm. It is structurally the most adapted to the approximation $p_{Q,\rho}$ since on each hyper-rectangle of $\mathcal{D}_{Q,\rho}$, $p_{Q,\rho}$ has the same mass as p_Q but distributes it uniformly. Hence, the discrepancy in mass allocation will be confined to small hyper-rectangles whose sides are of size at most ρ.

We will also use two stronger distances: the total variation (TV) distance $d_{TV}(p_1, p_2) = \|p_1 - p_2\|_{L^1(\mathcal{X})}$, and the Hellinger distance $H(p_1, p_2) = \|\sqrt{p_1} - \sqrt{p_2}\|_{L^2(\mathcal{X})}$, which is particularly relevant for exponential models [7], and, in our paper, when using rank-1 PSD models (see Sec. 4.2). These distances will naturally appear in Sec. 4 to quantify the discrepancy between a given probability density and its approximation as a Gaussian PSD model. For more details on these distances, see Appendix A.2. Theorem 2 provides bounds on these distances between the target density $p_Q = f\mathbf{1}_Q/I(Q)$ and $p_{Q,\rho}$ as a function of ρ, and some Lipschitz constant (where $\mathrm{Lip}_\infty(g)$ denotes the Lipschitz constant of g for the norm $\|x\|_\infty = \sup |x_i|$). A more general theorem is proved in Appendix D.3 as Theorem 7.

Theorem 2 (Variation bounds)**.** *Let Q be a hyper-rectangle, $\rho > 0$, $p_Q = f\mathbf{1}_Q/I(Q)$ and $p_{Q,\rho}$ defined in Eq. (7). It holds:*

$$H(p_Q, p_{Q,\rho}) \leqslant \sqrt{\tfrac{|Q|}{I(Q)}}\, \mathrm{Lip}_\infty(\sqrt{f})\, \rho \tag{9}$$

$$d_{TV}(p_Q, p_{Q,\rho}) \leqslant \tfrac{|Q|}{I(Q)}\, \mathrm{Lip}_\infty(f)\rho \tag{10}$$

$$\mathbb{W}_1(p_Q, p_{Q,\rho}) \leqslant \sqrt{d}\rho. \tag{11}$$

Combining the result of Theorems 1 and 2, we have that, given a PSD model on m centers, an hyper-rectangle of interest Q and an error ρ, algorithm 1 provides N i.i.d.samples whose distribution is distant $\sqrt{d}\rho$ in terms of $\mathbb{W}_1$ from the density represented by the PSD model over the hyper-rectangle. In particular, algorithm 1 computes the N i.i.d. samples with a cost of $O(N\, m^2\, d\, (\log_2(2|Q|) + d\log_2(2/\rho)))$.

Selection of ρ. In Fig. 1, we observe the effect of ρ on the quality of sampling, when sampling from a PSD model whose distribution is illustrated by the heat map defined on the top left figure. We highlight the fact that decreasing ρ corresponds to refining the dyadic decomposition of the hyper-rectangle and hence sampling more precisely. In practice, one can therefore choose ρ manually (for instance $\rho = 10^{-4}, 10^{-6}$) and have an upper bound on the distance between $p_{Q,\rho}$ and p_Q from Theorem 2. If one wishes to select ρ in a more principled way to bound the total variation or Hellinger distance, this can also be done using only accessible quantities. If f is a PSD model with parameters (A, X, η) for $\eta = \tau\mathbf{1}_d$, and K is a shorthand for $K_{X,\eta}$, the Lipschitz constants can be bounded using only τ, K and A (or a s.t. $A = aa^\top$ in the case of a rank one PSD model). More precisely, it holds

$$\mathrm{Lip}_\infty(f) \leqslant \sqrt{8\tau}d\|K^{1/2}AK^{1/2}\| =: \widetilde{\mathrm{Lip}}(A) \tag{12}$$

$$\mathrm{Lip}_\infty(\sqrt{f}) \leqslant \sqrt{2\tau}d\|K^{1/2}a\| =: \widetilde{\mathrm{Lip}}(a), \tag{13}$$

where for Eq. (13), $A = aa^\top$ is assumed to be a rank-1 operator[1]. These quantities only depend on a, A, K and can be computed explicitly. Combining these bounds with Eqs. (9) and (10), ρ can be selected in an adaptive way in algorithm 1.

[1]See Lemma 5 in Appendix C.1 for a proof of a the bound on $\mathrm{Lip}_\infty(f)$ when f is a PSD model and Lemma 2 in Appendix B.1 for a proof of of a bound on $\mathrm{Lip}_\infty(\sqrt{f})$ in the case where f is a rank one PSD model.

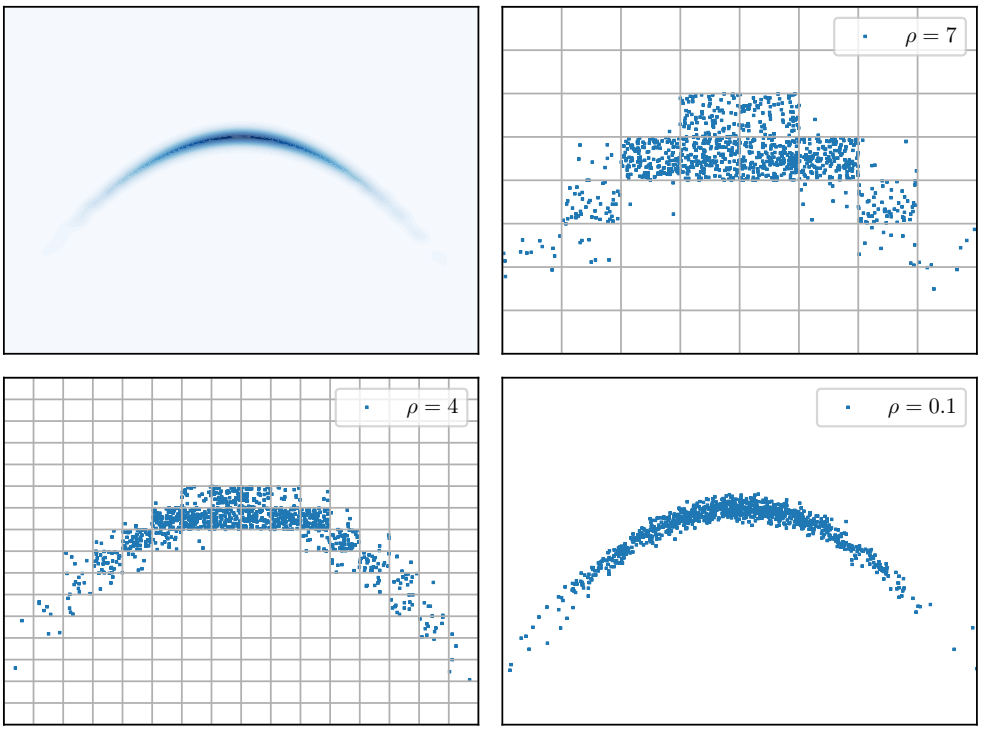

Figure 1: Samples obtained from algorithm 1 using different values for ρ

Remark 1 (Adaptive selection of ρ)**.** *Let $\varepsilon > 0$. Let f be a PSD model with matrix of coefficients A. Define*

$$\rho_\varepsilon^{TV} = \frac{I(Q)\varepsilon}{|Q|\widetilde{\mathrm{Lip}}(A)}, \qquad \rho_\varepsilon^{H} = \frac{\sqrt{I(Q)}\varepsilon}{\sqrt{|Q|\widetilde{\mathrm{Lip}}(a)}}, \tag{14}$$

where ρ_ε^H is defined if $A = aa^\top$ is a rank one matrix. If $\rho = \rho_\varepsilon^{TV}$ (resp. $\rho = \rho_\varepsilon^H$), then algorithm 1 applied to (Q, N, ρ) returns N i.i.d. samples from a distribution $p_{Q,\varepsilon}$ which satisfies $d_{TV}(p_Q, p_{Q,\varepsilon}) \leqslant \varepsilon$ (resp. $H(p_Q, p_{Q,\varepsilon}) \leqslant \varepsilon$).

3.2 Discussion

Sampling from the distribution on $\mathbb{R}^d$. It is possible to approximately sample from an infinite hyper-rectangle. To do so, one has to find a large enough hyper-rectangle Q such that almost all the mass is contained on Q and then apply the previous algorithm to this hyper-rectangle. One can, for instance, use algorithm 2.

Algorithm 2 Finding an approximate support Q

function FindApproximateSupport($f(\cdot; A, X, \eta)$, δ)
  $Q = \prod_{1 \leqslant k \leqslant d} [\min_{1 \leqslant i \leqslant n} X_{ik}, \max_{1 \leqslant i \leqslant n} X_{ik}]$
  $I = I(\mathbb{R}^d)$
  while $I(Q)/I \leqslant 1 - \varepsilon$ **do**
    $Q =$ DoubleSize(Q)
  end while
end function

Note that one can also concentrate f a priori using only its parameters (X, A, η), using Eq. (56) of Lemma 4 in Appendix C.1. One can use this bound to bound the number of steps in

algorithm 2.

Generality of the algorithm. algorithm 1 only relies on the fact that one can compute integrals on hyper-cubes of the model f. If we were to replace the Gaussian kernel k_η by a kernel k, and therefore have a PSD model of the form $\sum_{ij} A_{ij} k(x, x_i) k(x, x_j)$ with another positive definite kernel and $A \in \mathbb{S}_+^m$, then one would be able to run the algorithm as soon as computations of the form $\int_Q k(x, x_i) k(x, x_j) dx$ were tractable. This would extend this framework to more general PSD models, described by Marteau-Ferey et al. [9].

4 Sampling from arbitrary distributions using PSD models

The previous section provides an algorithm to approximately sample from a distribution in the form of a PSD model. In this section, we show how to leverage that fact to be able to generate N approximate i.i.d. samples from a very general class of probability distributions on a hyper-rectangle $\mathcal{X} \subset \mathbb{R}^d$. The strategy is simple : a) approximate the target distribution p with a PSD model $\widehat{p}$, and b) approximately sample from the PSD model $\widehat{p}$ using the algorithm presented in Sec. 3. The main challenge is to quantify the distance between the target distribution p and its approximation $\widehat{p}$ as a PSD model.

Approaching a distribution by a PSD model by accessing the distribution through samples has been done in Sec. 3. of Rudi and Ciliberto [15]. Instead, in this work, we access the distribution through function evaluations, as our goal is to be able to generate samples. However, a similar algorithm can be implemented to learn a PSD model from function evaluations. Moreover, it can be analysed under the same conditions (see Assumption 1 and Sec. 4.1). This algorithm is based on the solving of a semi-definite program to find the matrix A to form a good approximation $f(x;\ A, \widetilde{X}_m, \eta)$ of the density p. In Sec. 4.2, we instead learn a rank-one PSD model, solving a least-squares problem (and not a semi-definite program) using tools from Rudi et al. [13, 14], Meanti et al. [10]. This algorithm, faster than the one based on the solving of a semi-definite program, requires a stronger assumption to be analysed, and is naturally adapted to densities of the form $p(x) \propto e^{-V(x)}$.

Main hyper-parameters. The two methods presented in this section (see Sec. 4.1 and Sec. 4.2) will have hyper-parameters $n, m, \tau, \lambda, \rho$.

The parameters n and m are integer; moreover, we will take two sequences of i.i.d. samples uniformly from $\mathcal{X}$: $x_1, ..., x_n$ represented by $X \in \mathbb{R}^{n \times d}$ and $\tilde{x}_1, ..., \tilde{x}_m$ represented by $\widetilde{X}_m \in \mathbb{R}^{m \times d}$. We will use an isotropic $\eta = \tau \mathbf{1}_d$ in the Gaussian linear and PSD models for a strictly positive τ. To simplify notation, take $K_{mm} := K_{\widetilde{X}_m, \eta}$ and $K_{nm} := K_{X, \widetilde{X}_m, \eta}$. The parameter λ will always be a strictly positive real number.

The parameters m and τ will define the PSD model: m will control the number of points, also called *Nyström centers*, which we use to represent our PSD model (as n and m increase, the quality of the approximation increases); and τ will control the width of the Gaussian kernel. The parameter n and λ control the learning phase of the algorithm, i.e., the approximation of p by a PSD model. n is the number of points at which we evaluate our probability density to estimate it; λ will control the strength of the regularization. Finally, ρ will control the scale at which we apply algorithm 1.

4.1 A general method

In this section, we present a method to approximately sample from the density by a) approximating it by a PSD model solving a semi-definite program (SDP) and b) use algorithm 1 to sample from

that PSD model. More precisely, we assume that p is known up to a constant, i.e., that we have a function f_p which is proportional to p which we can evaluate.

Step a): approximation of p. To fit a PSD model to p, we use an method similar to the one presented in Section 3 of Rudi and Ciliberto [15], and construct a Gaussian PSD model $\widehat{f} = f(\bullet;\ \widehat{A}, \widetilde{X}_m, \eta)$, where $\widehat{A} \in \mathbb{S}_+^m$ is the solution to the empirical semi-definite problem

$$\widehat{A} = \operatorname*{argmin}_{A \in \mathbb{S}_+^m} \int_{\mathcal{X}} f(x;\ A)^2 dx$$
$$-2 \sum_{i=1}^{n} f_p(x_i) f(x_i;\ A) + \lambda \| K_{mm}^{1/2} A K_{mm}^{1/2} \|_F^2, \tag{15}$$

where $f(x;\ A) := f(x;\ A, \widetilde{X}_m, \eta)$. This problem is a quadratic problem in A and can be solved in polynomial time in m using semi-definite programming. We then define $\widehat{Z} = \int_{\mathcal{X}} \widehat{f}(x)\, dx$ which can be computed in closed form as the integral over a hyper-cube of a PSD model, and $\widehat{p} = \widehat{f}/\widehat{Z}$, which is our approximation of p.

Problem Eq. (15) can be seen as a variation of empirical risk minimization for the square loss, with an additional regularization term $\lambda \| K_{mm}^{1/2} A K_{mm}^{1/2} \|_F$ which is the equivalent of the classical kernel regularization term in the setting of PSD models. Indeed, the function of A being minimized is a proxy of $\| f(\cdot;\ A) - f_p(\cdot) \|_{L^2(\mathcal{X})}^2 = \| f(\cdot;\ A) \|_{L^2(\mathcal{X})}^2 + \int_{\mathcal{X}} f_p(x) f(x;\ a)\, dx + C$. In Eq. (15), $\int_{\mathcal{X}} f_p(x) f(x;\ a)\, dx$ is approximated by its empirical version, using uniform samples $X = (x_1, ..., x_n)$ (plus the regularization term). The first term $\| f(\cdot;\ A) \|_{L^2(\mathcal{X})}^2$ is kept as such as it is a quadratic function of A which can be explicitly computed, using the same techniques as those to compute integrals of PSD models, and described by Rudi and Ciliberto [15]. Note that here, $X, \widetilde{X}_m, \tau, \lambda$ are hyper-parameters; n and m will be taken as large as possible with a given computational budget, and λ and τ can be selected by validation on a newly generated test data set (since we assume we can generate samples from $\mathcal{X}$).

Step b): sampling from the approximation $\widehat{p}$. We apply algorithm 1 to $\widehat{p}$ with a parameter ρ and on the hyper-rectangle $\mathcal{X}$. We denote with $p_{\texttt{sample}}$ the density $\widehat{p}_{\mathcal{X},\rho}$ given by Eq. (7), from which algorithm 1 effectively samples N i.i.d. samples by Theorem 1. This two step strategy is detailed in algorithm 3. SolveSDP simply solves Eq. (15).

Algorithm 3 Approximately sampling from p using a SDP

Input $p, \mathcal{X}, N$
Hyper-parameters (approximation) n, m, τ, λ
Hyper-parameters (sampling) ρ
Output N approximate samples from $p|_{\mathcal{X}}$

function ApproximateSamples($p, \mathcal{X}, N, n, m, \tau, \lambda, \rho$)
  X_n = UniformSamples($n, \mathcal{X}$)
  X_m =UniformSamples($m, \mathcal{X}$)
  A = SolveSDP($p, X_n, X_m, \tau, \lambda$)
  $\widehat{p}(\cdot) = f(\cdot \mid A, X_m, \tau)$
  X_N = Sample($\mathcal{X}, N, \rho$) from $\widehat{p}$
  return X_N
end function

Theoretical analysis. Recall that p is the target density, proportional to f_p and that $\widehat{p}$ is the approximation of p obtained by solving Eq. (15) and $p_{\texttt{sample}}$ is the distribution from which we effectively sample when applying algorithm 1 to $\widehat{p}$. In proposition 1 and Theorem 3, we show that under certain regularity assumptions on p, given $\varepsilon > 0$, we can find hyper-parameters n, m, τ, λ and ρ such that $d_{TV}(p, p_{\texttt{sample}}) \leqslant C\varepsilon$, i.e. that algorithm 3 generates N i.i.d. samples from a distribution $C\varepsilon$ close to p.

For simplicity, we will assume $\mathcal{X} = (-1, 1)^d$, as is done by Rudi and Ciliberto [15]. In principle, we could approximate p on any bounded domain $\mathcal{X}$ from which we can sample uniformly, and still obtain analogous results. In that case, we would apply algorithm 1 on a hyper-rectangle containing the domain, and reject a sample outside of it. Our main assumption on p will be that p can be written as a sum of squares of functions belonging to the space $\widetilde{W}^\beta(\mathcal{X}) = W_2^\beta(\mathcal{X}) \cap L^\infty(\mathcal{X})$ which is the space of bounded functions whose derivatives of order less or equal to β are square integrable, and which can be equipped with the norm $\|\cdot\|_{\widetilde{W}^\beta(\mathcal{X})} = \|\cdot\|_{W_2^\beta(\mathcal{X})} + \|\cdot\|_{L^\infty(\mathcal{X})}$ (see Appendix A.1 for more precise definitions). The key quantities here are the dimension d and the regularity of the density β. This summarized in the following assumption.

Assumption 1 (Sum of squares distribution)**.** *There exists $J \in \mathbb{N}$ and functions $q_1, ..., q_J$ belonging to $\widetilde{W}^\beta(\mathcal{X})$ such that $p = \sum_{j=1}^J q_j^2$. Moreover, we have access to p only through function evaluations of the form $f_p(x)$ where $f_p \geqslant 0$ is given, is proportional to p, and where the proportionality constant is unknown. We define $\|p\|_{\texttt{sos},\mathcal{X},\beta} = \inf \sum_{j=1}^J \|q_j\|^2_{\widetilde{W}^\beta(\mathcal{X})}$ where the infimum is taken over all such decompositions of p.*

The approximation properties of $\widehat{p}$ w.r.t. p are bounded in total variation distance in the following proposition, proved as proposition 10 in Appendix E.

Proposition 1 (Performance of $\widehat{p}$)**.** *There exist constants $\varepsilon_0 > 0$ depending only on d, β, and $\|p\|_{\texttt{sos},\mathcal{X},\beta}$ and C_1, C_1', C_2', C_3' depending only on d, β such that the following holds. Let $\delta \in (0, 1]$ and $\varepsilon \leqslant \varepsilon_0$, and assume n and m satisfy*

$$m \geqslant C_1' \varepsilon^{-d/\beta} \log^d\left(\frac{C_2'}{\varepsilon}\right) \log\left(\frac{C_3'}{\varepsilon\delta}\right), \tag{16}$$

$$n \geqslant \varepsilon^{-2-d/\beta} \log^d\left(\frac{1}{\varepsilon}\right) \log\left(\frac{2}{\delta}\right). \tag{17}$$

Let $\lambda = \varepsilon^{2+2d/\beta}$ and $\tau = \varepsilon^{-2/\beta}$. With probability at least $1 - 2\delta$, it holds

$$d_{TV}(\widehat{p}, p) \leqslant C_1 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon. \tag{18}$$

The key takeaway from this proposition is that the number of samples n, m needed to perform the first step of the algorithm (approximation) is polynomial in the quantities $O(\varepsilon^{-1}), O(\varepsilon^{-d/\beta})$, thus leveraging the regularity β of p. When this is the case, we can find λ, τ such that the distance $d(p, \widehat{p})$ is of order ε. We provide a choice for ρ for the second step of the algorithm (sampling), in order to guarantee a bound for the total variation distance between the sampling distribution and the original distribution in the following theorem. It is proved as Theorem 8 in Appendix E. In particular, it bounds the total complexity of the algorithm in terms of erf computations, as a function of N and the desired error ε.

Theorem 3 (Performance of $p_{\texttt{sample}}$)**.** *Under the assumptions and notations of proposition 1, there exists a constant C_2 depending only on d, β, such that the following holds. If ρ is set either as $\varepsilon^{1+(d+1)/\beta}$ or adaptively as ρ_ε^{TV}, then with probability at least $1 - 2\delta$,*

$$d_{TV}(p, p_{\texttt{sample}}) \leqslant C_2 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon. \tag{19}$$

Moreover, the adaptive ρ_ε^{TV} *is lower bounded by* $\varepsilon^{1+(d+1)/\beta}/(C_3\ \|p\|_{\mathsf{sos},\mathcal{X},\beta})$. *In both cases, this guarantees that the complexity in terms of* erf *computations is of order* $O(Nm^2\log(1/\rho))$, *which in terms of* ε *yields* $O\left(N\,\varepsilon^{-2d/\beta}\log^{2d+1}\left(\frac{1}{\varepsilon}\right)\log^2\left(\frac{1}{\delta\varepsilon}\right)\right)$, *where the* O *notations is taken with constants depending on* $d, \beta, \|p\|_{\mathsf{sos},\mathcal{X},\beta}$.

4.2 Efficient method with a rank one model

In this section, we present a method to approximately sample from the density p by approximating it by a PSD model solving a linear system (as opposed to a SDP). This simpler and faster method comes at the expense of the stronger Assumption 2 needed to provide guarantees. As for algorithm 3, we first approximate the density with a PSD model and then sample from it using algorithm 1. The difference lies in the approximation step. We assume that we can evaluate a function g_p such that $g_p^2 \propto p$ (usually, this function will be proportional to the square root of p). We then approximate g_p with a Gaussian linear model Eq. (2) by solving a regularized empirical least squares problem, which is much faster than the solving of a SDP. Taking the square of that linear model, we obtain a PSD approximation of p from which we can sample using algorithm 1.

Step a): approximation of p. To fit a PSD model to p, we start by approximating g_p by a linear model $\widehat{g} = g(\bullet;\ \widehat{a}, \widetilde{X}_m, \eta)$ (see Eq. (2)), where $\widehat{a} \in \mathbb{R}^m$ is the solution to the empirical problem

$$\min_{a\in\mathbb{R}^m} \frac{1}{n}\sum_{i=1}^{n} |g(x_i;\ a) - g_p(x_i)|^2 + \lambda a^\top K_{mm} a, \tag{20}$$

where $g(x;\ a) := g(x;\ a, \widetilde{X}_m, \eta)$ and $g_n = (g_p(x_i))_{1\leqslant i\leqslant n}$. $\widehat{a}$ is the solution to the system :

$$\left(K_{nm}^\top K_{nm} + (\lambda n) K_{mm}\right) a = K_{nm}^\top g_n, \tag{21}$$

which can be solved either directly in time $O(nm^2 + m^3)$ [13] or using a pre-conditioned conjugate gradient method in time $O(m^3 + nm)$ [14, 10, 8]. We then define $\widehat{f} = \widehat{g}^2$ which is a rank-1 PSD model with coefficients $\widehat{A} = \widehat{a}\widehat{a}^\top$, $\widehat{Z} = \int_{\mathcal{X}} \widehat{f}(x)dx = \|\widehat{g}\|^2_{L^2(\mathcal{X})}$ which is computable in closed form as the integral of a PSD model (see Eq. (3)), and our approximation $\widehat{p} = \widehat{f}/\widehat{Z}$ of p.

Solving Eq. (20) can be seen as solving a regularized empirical risk minimization problem for the Hellinger distance (see Eq. (32) in Sec. 3.1); the regularization term $\lambda a^\top K_{mm} a$ being a regularization in the norm of the RKHS associated to the Gaussian kernel (see Appendix B). The Hellinger distance is particularly adapted to exponential models of the form $\exp(-V(x))$ for a real-valued potential V, as the square root is simply $\exp(-V(x)/2)$.

Step b): sampling from the approximation $\widehat{p}$. We apply algorithm 1 to $\widehat{p}$ with a parameter ρ and on the hyper-rectangle $\mathcal{X}$. We denote with $p_{\texttt{sample}}$ the density $\widehat{p}_{\mathcal{X},\rho}$ given by Eq. (7), from which algorithm 1 effectively samples N i.i.d. samples by Theorem 1. This two step strategy is detailed in algorithm 4. SolveHellinger simply solves Eq. (20).

Theoretical Analysis We use the same notation as introduced in Sec. 4.1. Once again, we assume that $\mathcal{X} = (-1, 1)^d$ for simplicity. In order to obtain good learning rates for algorithm 4, we make the following assumption, which is stronger than Assumption 1: it assumes that p can be written as a *single* square q^2, where q belongs to $\widetilde{W}^\beta(\mathcal{X})$.

Algorithm 4 Sampling from p using a rank-1 model

Input $p, \mathcal{X}, N$
Hyper-parameters (approximation) n, m, τ, λ
Hyper-parameters (sampling) ρ
Output N approximate samples from $p|_{\mathcal{X}}$

function ApproximateSamples($p, \mathcal{X}, N, n, m, \tau, \lambda, \rho$)
  X_n = UniformSamples($n, \mathcal{X}$)
  X_m =UniformSamples($m, \mathcal{X}$)
  A = SolveHellinger($p, X_n, X_m, \tau, \lambda$)
  $\widehat{p}(\cdot) = f(\cdot \mid A, X_m, \tau)$
  X_N = Sample($\mathcal{X}, N, \rho$) from $\widehat{p}$
  return X_N
end function

Assumption 2 (Square distribution)**.** *There exists a function q belonging to $\widetilde{W}^{\beta}(\mathcal{X})$ such that $p = q^2$. Moreover, we have access to p only through function evaluations of the form $g_p(x)$, where $g_p \propto q$ and where the proportionality constant is unknown.*

Note that this assumption is satisfied if $p \propto e^{-V(x)}$ for a potential V which is β times continuously differentiable which we can evaluate.

In proposition 2 and Theorem 4, we show that under certain regularity assumptions on p, given $\varepsilon > 0$, we can find hyper-parameters n, m, τ, λ and ρ such that $H(p, p_{\texttt{sample}}) \leqslant C\varepsilon$, i.e., that algorithm 4 generates N i.i.d. samples from a distribution $C\varepsilon$ close to p.

Proposition 2 (Performance of $\widehat{p}$)**.** *Let $\widetilde{\nu} > \min(1, d/(2\beta))$. There exists a constant ε_0 depending only on $\|q\|_{\widetilde{W}^{\beta}(\mathcal{X})}, \beta, d$, constants C_1, C_2, C_3, C_4 depending only on β, d and a constant C_1' depending only on $\beta, d, \widetilde{\nu}$ such that the following holds.*

Let $\delta \in (0, 1]$ and $\varepsilon \leqslant \varepsilon_0$, and assume m and n satisfy

$$m \geqslant C_1 \varepsilon^{-d/\beta} \log^d\left(\tfrac{C_2}{\varepsilon}\right) \log \tfrac{C_3}{\delta\varepsilon} \tag{22}$$

$$n \geqslant C_1' \varepsilon^{-2\widetilde{\nu}} \log \tfrac{8}{\delta}. \tag{23}$$

Let $\tau = \varepsilon^{-2/\beta}$ and $\lambda = \varepsilon^{2+d/\beta}$. With probability at least $1 - 3\delta$, it holds

$$H(\widehat{p}, p) \leqslant C_4 \|q\|_{\widetilde{W}^{\beta}(\mathcal{X})}\, \varepsilon. \tag{24}$$

Once again, the key takeaway from this proposition is that the number of samples n, m needed to perform the first step of the algorithm (approximation) is polynomial in the quantities $O(\varepsilon^{-1}), O(\varepsilon^{-d/\beta})$, thus leveraging the regularity β of q s.t. $q^2 = p$. When this is the case, we can find λ, τ such that the distance $H(p, \widehat{p})$ is of order ε. We provide a choice for ρ for the second step of the algorithm (sampling), in order to guarantee a bound for the Hellinger distance between the sampling distribution and the original distribution in the following theorem. It is proved as Theorem 10 in Appendix F. In particular, it bounds the total complexity of the algorithm in terms of erf computations, as a function of N and the desired error ε.

Theorem 4 (Performance of $p_{\texttt{sample}}$)**.** *Under the assumptions and notations of proposition 2, there exists a constant C_5 depending only on d, β, such that the following holds. If on the one hand ρ is set either as $\varepsilon^{1+(d+2)/(2\beta)}$ or adaptively as ρ_{ε}^{H} (see Remark 1), then with probability at least $1 - 3\delta$,*

$$H(p, p_{\texttt{sample}}) \leqslant C_5 \|q\|_{\widetilde{W}^{\beta}(\mathcal{X})}\, \varepsilon. \tag{25}$$

Moreover, the adaptive ρ_ε^H *is lower bounded by* $\varepsilon^{1+(d+2)/\beta}/(C_5\ \|q\|_{\widetilde{W}^\beta(\mathcal{X})})$. *In both cases, this guarantees that the complexity in terms of* erf *computations is bounded by* $O(Nm^2 \log \frac{1}{\rho})$, *which, in terms of* ε, *yields* $O\left(N\ \varepsilon^{-2d/\beta} \log^{2d+1}\left(\frac{1}{\varepsilon}\right) \log^2\left(\frac{1}{\delta\varepsilon}\right)\right)$ *where the* O *notation incorporates constants depending on* $d, \beta, \|q\|_{\widetilde{W}^\beta(\mathcal{X})}$.

4.3 Discussion

The two methods presented in Sec. 4.1 and Sec. 4.2 share many interesting properties, both from a practical and theoretical viewpoint.

On the theoretical side, even though we only have access to the distribution up to a re-normalizing constant, this does not influence the theoretical results, i.e., the bounds we get only depend on the density p through its norm $\|p\|$. Moreover, the number of samples n, m needed (and hence the complexity of the sampling and of the approximation algorithm) is polynomial in the quantities $O(\varepsilon^{-1}), O(\varepsilon^{-d/\beta})$, showing that as soon as $\beta \geqslant d$, the dimension plays no role in the exponents of these error terms and thus *breaking the curse of dimensionality* in the rates. However, the constants in the $O(\cdot)$ term can be exponential in d, and without more hypotheses, **they are unimprovable** [11]. We therefore keep a form of "curse of dimensionality" in the constants, but not in the rate. Concretely this means that we need a number of points in the order of the constants before having a reasonable error (i.e., $\varepsilon = 1$). However, as soon as this number is reached, one can rapidly gain in precision, if the function is regular. Moreover, in practice, we do not always pay this exponential constant, owing to some additional regularity of the function. Interestingly, this phenomenon is shared with approximation, learning and optimization problems over a wide family of functions (see [11] for more details).

On the practical side, note that both algorithm 3 and algorithm 4 can be run for any hyper-parameter (even though this might not have statistical sense), making it easy to use. More importantly, we can evaluate the learnt model a posteriori using empirical metrics (like the empirical total variation distance or the empirical Hellinger distance for instance) on a new data set generated uniformly from $\mathcal{X}$. We could also evaluate it using certain empirical divergences since we are able to sample from $p_{\texttt{sample}}$. This can help in both selecting τ and λ by validation, as well as in simply evaluating the performance of the learnt model, with error bars if needed. In Fig. 2 for example, we evaluate the performance of learnt PSD models for the empirical Hellinger distance. We perform 5 different tests and plot the associated error bars: this methods seems very robust for evaluation.

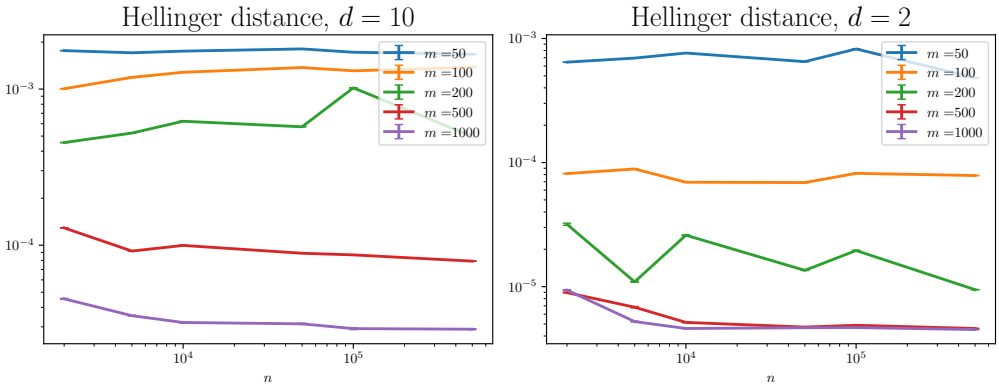

Figure 2: Evolution of the empirical Hellinger distance on a test set, between learnt distribution $\widehat{p}$ and target distribution p when increasing the number of evaluation points, for fixed values of m. We learn $\widehat{p}$ as a rank one PSD model through Eq. (20). *(left)* Learning p_2 with $d = 10$ defined in Sec. 5. *(right)* Learning p_1 defined in Sec. 5.

5 EXPERIMENTS

The experiments in this work were executed on a MacBook Pro equipped with a 2.8 GHz Quad-Core Intel Core i7 processor and 16Gb of RAM[2].

Influence of m and n. In Fig. 2, we show how m and n interact in order to set the precision of our approximation in the learning phase (step a)). For $m = 50, 100, 200$, m is so small that increasing n beyond 1000 does not yield better performance (the variations are due to the fact that points are always resampled accross experiments). However, when $m = 500, 1000$, we see that increasing n yields better performance, before arriving at a plateau. This plateau corresponds to the transition from the phase where n is the limiting statistical factor to the phase where m is.

Qualitative performance of our algorithm. In Fig. 3, we show an example of the way our algorithm approximates a certain target density p_1 known up to a renormalization constant: $p_1(x) \propto 0.08k_{0.7}(x, -(1,1)) - 0.4k_{0.6}(x,(1,1)) + 0.4k_{0.7}(x,(1,1))$. In the top left figure, a heat map of p_1 is plotted. Note that p_1 is not a Gaussian PSD model, as the widths of the Gaussian kernels are not the same. We then use algorithm 4 to approximate p_1 by a rank one PSD model $\widehat{p}_1$ (whose heat-map is plotted on the top right figure) and then sample $N = 1000$ samples from this approximation (plotted in the bottom left figure). Note that in order to approximate p_1 by $\widehat{p}_1$, $n = 10^5, m = 300$ were fixed and $\tau = 2, \lambda = 10^{-9}$ were selected on a test set. In Appendix G, we perform and comment another experiment when trying to learn a density which is not smooth (and therefore out of the scope of Theorems 3 and 4).

Quantitative performance of our algorithm. To further demonstrate the promising nature of our sampling algorithm, we tried learning the density $p_2(x) \propto (k_{1/5}(x,(1,...,1)) - k_{1/5}(x,-(1,...,1)))^2$ on $Q = [-1,1]^d$, for $d = 5$. Contrary to p_1, this is a PSD model, we can sample from it with very high precision (here, we chose $\rho = 10^{-6}$). Our goal here is to be able to compare methods through the generated samples.

We compared the performance of our model to the naive gridding algorithm which, if allowed n function evaluations, computes a grid G of side $n^{1/d}$, which we identify to the set of centers of the tiles of the grid, and evaluates p at each point in the grid. To sample a point, one chooses a point $g \in G$ with probability $p(g)/\sum_{h\in G} p(h)$, and then draws a sample uniformly in that tile. It is the algorithm called 'grid' in the bottom right figure of Fig. 3.

We compare our algorithm with the gridding algorithm by fixing the number n of function evaluations of p each method is allowed, and computing the distance between each method and the ground truth. The distance we use between distributions is the empirical version of the Maxmium Mean Discrepancy distance (MMD) [20, 19], which is defined, for the Gaussian kernel k_η of parameter η, as $d_\eta(p,\tilde{p}) = \|\mathbb{E}_{X\sim p}[\phi_\eta(X)] - \mathbb{E}_{X\sim\tilde{p}}[\phi_\eta(X)]\|_{\mathcal{H}_\eta}$ where ϕ_η is the embedding associated to the Gaussian kernel k_η (for more details, see Appendix A). This distance can be approximated using N samples $(x_i)_{1\leqslant i\leqslant N}$ from p and N samples $(\tilde{x}_j)_{1\leqslant j\leqslant N}$ from $\tilde{p}$ as $\widehat{d}_\eta(p,\tilde{p}) = \left\| \frac{1}{N}\sum_{i=1}^N \phi_\eta(x_i) - \frac{1}{N}\sum_{j=1}^N \phi_\eta(\tilde{x}_j) \right\|_{\mathcal{H}_\eta}$. This quantity can be computed explicitly using kernel matrices [20]. However, Tolstikhin et al. [22] show that the minimax rate cannot exceed $1/\sqrt{N}$, i.e., that $\widehat{d}_\eta$ approximates d_η only with precision of order $1/\sqrt{N}$.

In our experiments, we take $N = 10^4$. We compute the empirical distances $\widehat{d}_\eta$ five times using newly generated samples from each distribution, and compute an empirical mean and standard

[2]The code is available at https://github.com/umarteau/sampling_psd_models

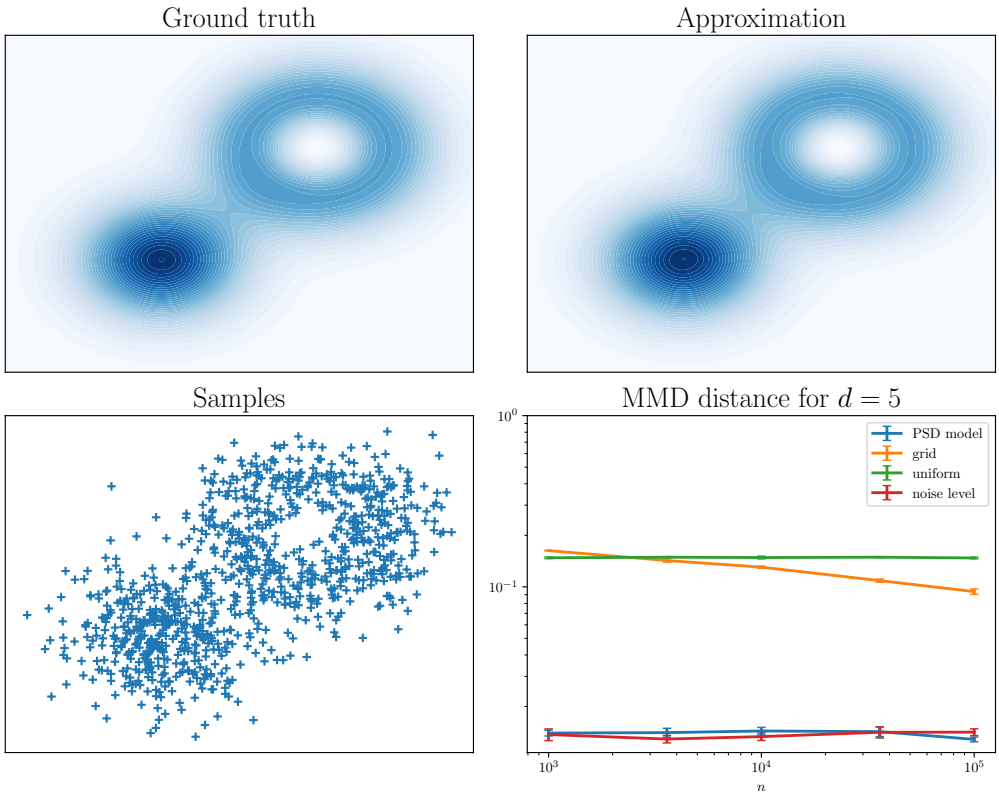

Figure 3: *(top left)* Plot of the distribution p_1, *(top right)* heat map of an approximation $\widehat{p}_1$ of p_1. *(bottom left)* samples generated from $\widehat{p}_1$, *(bottom right)* performance of our method in MMD distance.

deviation, reported as error bars on the plot. When approximating p_2 by a PSD model using algorithm 4, we take $m = 50$, as there is no need to increase m to reach better precision than the target distribution for $\widehat{d}_\eta$. We take $\rho = 10^{-3}$ and select τ, λ by using half of the evaluation points as a test set.

The results reported on the bottom-right plot of Fig. 3 show that in dimension 5, the 'grid' method is not competitive anymore, and is close to the uniform distribution in performance for $\eta = 2$. Note that the choice of η in a wide range from 0.1 to 10 does not change these results. They also show that when taking only $N = 10^4$ to approximate the MMD distance, our method is below the noise level.

6 Extensions, future work

In this paper, we have introduced a method for sampling any distribution from function values by first approximating it with a so-called PSD model and then sampling from this PSD model using the algorithm introduced in Sec. 3.

Natural extensions of this work include the fact that while we cast a least squares problem in Sec. 4.1, we can actually minimize more general convex losses adapted to distributions, such as maximum log-likelihood estimation. Moreover, as mentioned in Sec. 3, the proposed algorithm only relies on integral computations, and could therefore be extended to other kernels, provided they can easily be integrated on hyper-rectangles.

Future work will start with trying to scale the sampling method up in terms of generation of samples,

by both theoretical means (to make computation saving approximations) and computational means (use of GPUs, parallelization).

Acknowledgements. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001(PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grants SEQUOIA 724063 and REAL 947908), and support by grants from Région Ile-de-France.

Bibliography

[1] Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*. Elsevier.

[2] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

[3] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: a non Asymptotic Theory of Independence*. Oxford University Press.

[4] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition.

[5] Lelièvre, T., Rousset, M., and Stoltz, G. (2010). *Free Energy Computations*. IMPERIAL COLLEGE PRESS.

[6] Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer Publishing Company, Incorporated.

[7] Lucien Le Cam, G. L. Y. a. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics. Springer US.

[8] Marteau-Ferey, U., Bach, F., and Rudi, A. (2019). Globally convergent Newton methods for ill-conditioned generalized self-concordant losses. In *Advances in Neural Information Processing Systems*, volume 32.

[9] Marteau-Ferey, U., Bach, F., and Rudi, A. (2020). Non-parametric models for non-negative functions. *Advances in Neural Information Processing Systems*, 33.

[10] Meanti, G., Carratino, L., Rosasco, L., and Rudi, A. (2020). Kernel methods through the roof: Handling billions of points efficiently. In *Advances in Neural Information Processing Systems*, volume 33, pages 14410–14422.

[11] Novak, E. (2006). *Deterministic and Stochastic Error Bounds in Numerical Analysis*, volume 1349. Springer.

[12] Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.

[13] Rudi, A., Camoriano, R., and Rosasco, L. (2015). Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665.

[14] Rudi, A., Carratino, L., and Rosasco, L. (2017). Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3888–3898.

[15] Rudi, A. and Ciliberto, C. (2021). Psd representations for effective probability models. *Advances in Neural Information Processing Systems*, 34.

[16] Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with random features. *Advances in Neural Information Processing Systems*, 30:3215–3225.

[17] Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Springer International Publishing.

[18] Schölkopf, B. and Smola, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

[19] Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410.

[20] Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561.

[21] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media.

[22] Tolstikhin, I. O., Sriperumbudur, B. K., and Schölkopf, B. (2016). Minimax estimation of maximum mean discrepancy with radial kernels. In *Advances in Neural Information Processing Systems*, volume 29.

[23] Vacher, A., Muzellec, B., Rudi, A., Bach, F., and Vialard, F.-X. (2021). A dimension-free computational upper-bound for smooth optimal transport estimation. In *Conference on Computational Learning Theory*.

[24] Wendland, H. (2004). *Scattered Data Approximation*, volume 17. Cambridge University Press.

Organization of the Supplementary Material

A. Definitions and notations

We set the main notations and tools of the appendix (Fourier transform, vector and matrix notations, notations concerning hyper-rectangles, RKHS and specifically the Gaussian kernel).

A.1. Sobolev spaces

In this section, we focus more on notations and basic results concerning Sobolev spaces, as they will be our main tool to measure the regularity of a function.

A.2. Measuring distances between probability densities

In this section, we define and compare the basic distances we will be using to compare probability distributions in the paper, since we are always "approximating" a certain distribution with another. In particular, we define the total variation, Hellinger and Wasserstein distances.

A.3. General PSD models

We define PSD models in general [9, 15]. They will be our main tool for approximation and sampling, and relates to the more restrictive definition in Sec. 2.

B. Properties of the Gaussian RKHS

Throughout the paper the Gaussian kernel k_η and the associated Gaussian RKHS will be central objects. We introduce different properties and results.

B.1. Properties of the Gaussian kernel k_η

We introduce certain properties of the Gaussian kernel involving products, as well as a bound on the derivative of the associated embedding in Lemma 2.

B.2. Useful Matrices and Linear Operators on the Gaussian RKHS

We introduce the most important theoretical objects of the paper. We introduce kernel matrices, matrices which will appear in the integration of Gaussian PSD models, operators which relate L^2 to the RKHS $\mathcal{H}_\eta$, operators which allow to discretize using samples and "compression" operators which allow concise representations.

B.3. Approximation properties of the Gaussian kernel

We prove two important results concerning the approximation properties of the Gaussian RKHS in proposition 7 and the concise representation of models in Lemma 3.

C. Properties of Gaussian PSD models

We present the results specific to Gaussian PSD models. These results are often reformulations of theorems presented by Rudi and Ciliberto [15].

C.1. Bounds on the support and the derivatives

We present result to understand how the mass of a Gaussian PSD model is concentrated (Lemma 4) and how the derivative of a Gaussian PSD model can be bounded using only its parameters (Lemma 5).

C.2. Compression as a Gaussian PSD model

We restate Theorem C.4 of Rudi and Ciliberto [15] as Theorem 5 on the effect of a compression operator on a PSD model.

A Definitions and notations

In this section we recall results from Rudi and Ciliberto [15] which will be useful in the different statements and proofs.

Basic vector and matrix notations. Let $n, d \in \mathbb{N}$. We denote by $\mathbb{R}^d_{++}$ the space vectors in $\mathbb{R}^d$ with positive entries, $\mathbb{R}^{n\times d}$ the space of $n \times d$ matrices, $\mathbb{S}^n_+ = \mathbb{S}_+(\mathbb{R}^n)$ the space of positive semidefinite $n \times n$ matrices. Given a vector $\eta \in \mathbb{R}^d$, we denote $\mathrm{diag}(\eta) \in \mathbb{R}^{d\times d}$ the diagonal matrix associated to η. We denote by $A \circ B$ the entry-wise product between two matrices A and B. We denote by $\|A\|_F, \|A\|, \det(A), \mathrm{vec}(A)$ and $A^\top$ respectively the Frobenius norm, the operator norm (i.e. maximum singular value), the determinant, the (column-wise) vectorization of a matrix and the (conjugate) transpose of A. With some abuse of notation, where clear from context we write element-wise products and division of vectors $u, v \in \mathbb{R}^d$ as $uv, u/v$. The term $\mathbf{1}_n \in \mathbb{R}^n$ denotes the vector with all entries equal to 1.

Hyper-rectangles Define a hyper-rectangle Q as a product of the form $\prod_{k=1}^d [a_k, b_k[$, where $a \leqslant b$. Given a hyper-rectangle Q we denote its extremities with $a(Q) \leqslant b(Q) \in \mathbb{R}^d$ (i.e. $Q = \prod_{k=1}^d [a_k(Q), b_k(Q)[$), and its side-lengths $\rho(Q) = b(Q) - a(Q)$. We sometimes omit Q when it is implied by the context.

We will also use the so-called *error function*, which is defined as follows :

$$\mathrm{erf}(x) = \tfrac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

This function is implemented as an elementary function in most libraries.

Multi-index notation Let $\alpha \in \mathbb{N}^d$, $x \in \mathbb{R}^d$ and f be an infinitely differentiable function on $\mathbb{R}^d$, we introduce the following notation

$$|\alpha| = \sum_{j=1}^d \alpha_i, \quad \alpha! = \prod_{j=1}^d \alpha_j!, \quad x^\alpha = \prod_{j=1}^d x_j^{\alpha_j}, \quad \partial^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

We introduce also the notation D^α that corresponds to the multivariate distributional derivative of order α and such that

$$D^\alpha f = \partial^\alpha f$$

for functions that are differentiable at least $|\alpha|$ times [1].

Fourier Transform Given two functions $f, g : \mathbb{R}^d \to \mathbb{R}$ on some set $\mathbb{R}^d$, we denote by $f \cdot g$ the function corresponding to *pointwise product* of f, g, i.e.,

$$(f \cdot g)(x) = f(x)g(x), \quad \forall x \in \mathbb{R}^d.$$

Let $f, g \in L^1(\mathbb{R}^d)$ we denote the *convolution* by $f \star g$

$$(f \star g)(x) = \int_{\mathbb{R}^d} f(y)g(x-y)dy.$$

We now recall some basic properties, that will be used in the rest of the appendix.

Proposition 3 (Basic properties of the Fourier transform [24], Chapter 5.2.)**.**

(a) *There exists a linear isometry* $\mathcal{F} : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ *satisfying*

$$\mathcal{F}[f] = \int_{\mathbb{R}^d} e^{-2\pi i \omega^\top x} f(x)\, dx \quad \forall f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d),$$

where $i = \sqrt{-1}$. *The isometry is uniquely determined by the property in the equation above.*

(b) *Let* $f \in L^2(\mathbb{R}^d)$, *then* $\|\mathcal{F}[f]\|_{L^2(\mathbb{R}^d)} = \|f\|_{L^2(\mathbb{R}^d)}$.

(c) *Let* $f \in L^2(\mathbb{R}^d), r > 0$ *and define* $f_r(x) = f(\frac{x}{r}), \forall x \in \mathbb{R}^d$, *then* $\mathcal{F}[f_r](\omega) = r^d \mathcal{F}[f](r\omega)$.

(d) *Let* $f, g \in L^1(\mathbb{R}^d)$, *then* $\mathcal{F}[f \cdot g] = \mathcal{F}[f] \star \mathcal{F}[g]$.

(e) *Let* $\alpha \in \mathbb{N}^d$, $f, D^\alpha f \in L^2(\mathbb{R}^d)$, *then* $\mathcal{F}[D^\alpha f](\omega) = (2\pi i)^{|\alpha|} \omega^\alpha \mathcal{F}[f](\omega)$, $\forall \omega \in \mathbb{R}^d$.

(f) *Let* $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, *then* $\|\mathcal{F}[f]\|_{L^\infty(\mathbb{R}^d)} \leqslant \|f\|_{L^1(\mathbb{R}^d)}$.

(g) *Let* $f \in L^\infty(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, *then* $\|f\|_{L^\infty(\mathbb{R}^d)} \leqslant \|\mathcal{F}[f]\|_{L^1(\mathbb{R}^d)}$.

Reproducing kernel Hilbert spaces for translation invariant kernels. We now list some important facts about reproducing kernel Hilbert spaces in the case of translation invariant kernels on $\mathbb{R}^d$. For this paragraph, we refer to Steinwart and Christmann [21], Wendland [24]. For the general treatment of positive kernels and Reproducing kernel Hilbert spaces, see Aronszajn [2], Steinwart and Christmann [21]. Let $v : \mathbb{R}^d \to \mathbb{R}$ such that its Fourier transform $\mathcal{F}[v] \in L^1(\mathbb{R}^d)$ and satisfies $\mathcal{F}[v](\omega) \geqslant 0$ for all $\omega \in \mathbb{R}^d$. Then, the following hold.

(a) The function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined as $k(x, x') = v(x - x')$ for any $x, x' \in \mathbb{R}^d$ is a positive kernel and is called *translation invariant kernel*.

(b) The *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ and its norm $\|\cdot\|_{\mathcal{H}}$ are characterized by

$$\mathcal{H} = \{f \in L^2(\mathbb{R}^d) \mid \|f\|_{\mathcal{H}} < \infty\}, \quad \|f\|_{\mathcal{H}}^2 = \int_{\mathbb{R}^d} \frac{|\mathcal{F}[f](\omega)|^2}{\mathcal{F}[v](\omega)} d\omega, \tag{26}$$

(c) $\mathcal{H}$ is a separable Hilbert space, whose inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is characterized by

$$\langle f, g \rangle_{\mathcal{H}} = \int_{\mathbb{R}^d} \frac{\mathcal{F}[f](\omega)\overline{\mathcal{F}[g](\omega)}}{\mathcal{F}[v](\omega)} d\omega.$$

In the rest of the paper, when clear from the context we will simplify the notation of the inner product, by using $f^\top g$ for $f, g \in \mathcal{H}$, instead of the more cumbersome $\langle f, g \rangle_{\mathcal{H}}$.

(d) The feature map $\phi : \mathbb{R}^d \to \mathcal{H}$ is defined as $\phi(x) = k(x - \cdot) \in \mathcal{H}$ for any $x \in \mathbb{R}^d$.

(e) The functions in $\mathcal{H}$ have the *reproducing property*, i.e.,

$$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}, x \in \mathbb{R}^d,$$

in particular $k(x', x) = \langle \phi(x'), \phi(x) \rangle_{\mathcal{H}}$ for any $x', x \in \mathbb{R}^d$.

We now introduce the main tool of our analysis, the Gaussian RKHS, which will be further explored in Appendix C.

Example 1 (Gaussian Reproducing Kernel Hilbert Space)**.** *Let* $\eta \in \mathbb{R}^d_{++}$ *and* $k_\eta(x, x') = e^{-(x-x')^\top \operatorname{diag}(\eta)(x-x')}$, *for* $x, x' \in \mathbb{R}^d$ *be the Gaussian kernel with precision* η. *The function* k_η *is a translation invariant kernel, since* $k_\eta(x, x') = v(x - x')$ *with* $v(z) = e^{-\|D^{1/2} z\|^2}, D = \operatorname{diag}(\eta)$

and $\mathcal{F}[v](\omega) = c_\eta e^{-\pi^2\|D^{-1/2}\omega\|^2}$, $c_\eta = \pi^{d/2}\det(D)^{-1/2}$, *for* $\omega \in \mathbb{R}^d$ *is in* $L^1(\mathbb{R}^d)$ *and satisfies* $\mathcal{F}[v](\omega) \geqslant 0$ *for all* $\omega \in \mathbb{R}^d$. *The associated reproducing kernel Hilbert space* $\mathcal{H}_\eta$ *is defined according to Eq. (26), with norm*

$$\|f\|^2_{\mathcal{H}_\eta} \;=\; \frac{1}{c_\eta}\int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2\, e^{\pi^2\|D^{-1/2}\omega\|^2}\, d\omega, \qquad \forall f \in L^2(\mathbb{R}^d). \tag{27}$$

The inner product and the feature map ϕ_η *are defined as in the discussion above.*

A.1 Sobolev spaces

Let $\beta \in \mathbb{N}, p \in [1,\infty]$ and let $\Omega \subseteq \mathbb{R}^d$ be an open set. The set $L^p(\Omega)$ denotes the set of p-integrable functions on Ω for $p \in [1,\infty)$ and that of the essentially bounded on Ω when $p = \infty$. The set $W_p^\beta(\Omega)$ denotes the Sobolev space, i.e., the set of measurable functions with their distributional derivatives up to β-th order belonging to $L^p(\Omega)$,

$$W_p^\beta(\Omega) = \{f \in L^p(\Omega) \mid \|f\|_{W_p^\beta(\Omega)} < \infty\}, \quad \|f\|^p_{W_p^\beta(\Omega)} = \sum_{|\alpha|\leqslant\beta} \|D^\alpha f\|^p_{L^p(\Omega)}, \tag{28}$$

where D^α denotes the distributional derivative. In the case of $p = \infty$,

$$\|f\|_{W_\infty^\beta(\Omega)} = \max_{|\alpha|\leqslant\beta} \|D^\alpha f\|_{L^\infty(\Omega)}$$

We now recall some basic results about Sobolev spaces that are useful for the proofs in this paper. First we start by recalling the restriction properties of Sobolev spaces. Let $\Omega \subseteq \Omega' \subseteq \mathbb{R}^d$ be two open sets. Let $\beta \in \mathbb{N}$ and $p \in [1,\infty]$. By definition of the Sobolev norm above we have

$$\|g|_\Omega\|_{W_p^s(\Omega)} \leqslant \|g\|_{W_p^s(\Omega')},$$

and so $g|_\Omega \in W_p^s(\Omega)$ for any $g \in W_p^s(\Omega')$. Now we recall the extension properties of Sobolev spaces, which will allow us to consider the case

The formal definition of a set with Lipschitz boundary is provided by Adams and Fournier [1]. Note that if $\mathcal{X} = (-1,1)^d$, as will be the case later on for simplicity, then $\mathcal{X}$ is bounded and has Lipschitz boundary.

The following result shows that being in an intersection space allows to extend the function to the whole of $\mathbb{R}^d$. This will be useful in order to use the properties of translation invariant kernels in order to approximate functions which are a priori defined only on $\mathcal{X}$ but which we extend using this result.

Proposition 4 (Corollary A.3 of Rudi and Ciliberto [15])**.** *Let* $\mathcal{X} \subset \mathbb{R}^d$ *be a non-empty open set with Lipschitz boundary. Let* $\beta \in \mathbb{N}, p \in [1,\infty]$. *Then for any function* $f \in W_p^\beta(\mathcal{X}) \cap L^\infty(\mathcal{X})$ *there exists an extension* $\tilde{f}$ *on* $\mathbb{R}^d$, *i.e. a function* $\tilde{f} \in W_p^\beta(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ *such that*

$$f = \tilde{f}|_\mathcal{X} \text{ a.e. on } \mathcal{X}, \quad \|\tilde{f}\|_{L^\infty(\mathbb{R}^d)} \leqslant C\|f\|_{L^\infty(\mathcal{X})}, \quad \|\tilde{f}\|_{W_p^\beta(\mathbb{R}^d)} \leqslant C'\|f\|_{W_p^\beta(\mathcal{X})}.$$

The constant C *depends only on* $\mathcal{X}, d$, *and the constant* C' *only on* $\mathcal{X}, \beta, d, p$

The following proposition gives an idea of what these intersection spaces contain.

Proposition 5 (Proposition A.4 of Rudi and Ciliberto [15].)**.** *Let* $\mathcal{X}$ *be an open bounded set with Lipschitz boundary. Let* f *be a function that is* m *times differentiable on the closure of* $\mathcal{X}$. *Then there exists a function* $\tilde{f} \in W_p^m(\mathcal{X}) \cap L^\infty(\mathcal{X})$ *for any* $p \in [1,\infty]$, *such that* $\tilde{f} = f$ *on* $\mathcal{X}$.

The following proposition provides a useful characterization of the space $W_2^\beta(\mathbb{R}^d)$ in terms of Fourier transform; this will be particularly useful when approximating functions in $W_2^\beta(\mathbb{R}^d)$ by functions in a Gaussian RKHS $\mathcal{H}_\eta$ using the characterization of the norm in terms of Fourier transform for those kernels in Eq. (26).

Proposition 6 (Characterization of the Sobolev space $W_2^k(\mathbb{R}^d)$, Wendland [24], Proposition A.5 of Rudi and Ciliberto [15])**.** *Let $k \in \mathbb{N}$. The norm of the Sobolev space $\|\cdot\|_{W_2^k(\mathbb{R}^d)}$ is equivalent to the following norm*

$$\|f\|'^2_{W_2^k(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 \, (1 + \|\omega\|^2)^k \, d\omega, \quad \forall f \in L^2(\mathbb{R}^d)$$

and satisfies

$$\tfrac{1}{(2\pi)^{2k}} \|f\|^2_{W_2^k(\mathbb{R}^d)} \leqslant \|f\|'_{W_2^k(\mathbb{R}^d)} \leqslant 2^{2k} \|f\|^2_{W_2^k(\mathbb{R}^d)}, \quad \forall f \in L^2(\mathbb{R}^d) \tag{29}$$

Moreover, when $k > d/2$, then $W_2^k(\mathbb{R}^d)$ is a reproducing kernel Hilbert space.

A.2 Measuring distances between probability densities

In this work, since our aim is to approximate a probability distribution, we will often compare probability distributions, with different distances.

To simplify definitions, we will only consider distances between probability densities p_1, p_2 defined on a Borel subset $\mathcal{X}$ of $\mathbb{R}^d$ with respect to the Lebesgue measure. Note that while the total variation distance, the Hellinger distance and the Wasserstein distance do not actually depend on the choice of such a base measure and can be defined intrinsically, the L^2 distance cannot; that is why it is less appropriate from a statistical point of view. We consider it here because it is the natural distance in which we are able to solve Eq. (15).

The total variation (TV) or L^1 distance :

$$d_{TV}(p_1, p_2) := \|p_1 - p_2\|_{L^1(\mathcal{X})} = \int_{\mathcal{X}} |p_1(x) - p_2(x)| dx. \tag{30}$$

This distance can also be expressed using a dual formulation (see Chapter 3.2 of Lucien Le Cam [7]).

$$d_{TV}(p_1, p_2) = \sup_{|f| \leqslant 1} \left| \int_{\mathcal{X}} f(x)(p_1(x) - p_2(x)) \, dx \right| \tag{31}$$

The Hellinger distance : (this distance is particularly suitable in the case of exponential models; see Lucien Le Cam [7] and in particular Chapter 3).

$$H(p_1, p_2) := \|\sqrt{p_1} - \sqrt{p_2}\|_{L^2(\mathcal{X})} = \left(\int_{\mathcal{X}} |\sqrt{p_1}(x) - \sqrt{p_2}(x)|^2 \, dx \right)^{1/2} \tag{32}$$

The Wasserstein distance In the case where $\mathcal{X}$ is bounded (for simplicity), the p Wasserstein distance for $p \geqslant 1$ (see chapter 5 of Santambrogio [17]):

$$\mathbb{W}_p^p(p_1, p_2) = \inf_{\gamma \in \Pi(p_1, p_2)} \int_{\mathcal{X} \times \mathcal{X}} |x - y|^p \, d\gamma(x, y), \tag{33}$$

where $\Pi(p_1, p_2)$ is the set of all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals p_1 and p_2. Note that one has the following easier dual formulation when $p = 1$ (see the chapter on Kantorovich duality by Santambrogio [17]):

$$\mathbb{W}_1(p_1, p_2) = \sup_{f \in \mathrm{Lip}_1(\mathcal{X})} \int_{\mathcal{X}} f(x)(p_1(x) - p_2(x))dx, \tag{34}$$

where $\mathrm{Lip}_1(\mathcal{X})$ is the set of 1-Lipschitz functions on $\mathcal{X}$. Wasserstein distances capture the moving of mass; they are quite weak but are well-adapted to capture the behavior of our sampling algorithm which approximates probability densities on each hyper-rectangle.

The L^2 distance :

$$\|p_1 - p_2\|_{L^2(\mathcal{X})} = \left(\int_{\mathcal{X}} (p_1(x) - p_2(x))^2 \, dx \right)^{1/2} \tag{35}$$

Relating these difference distances . The following well known bounds exist between distances.

$$H^2(p_1, p_2) \leqslant d_{TV}(p_1, p_2) \leqslant \sqrt{2} H(p_1, p_2). \tag{36}$$

Moreover, if $\mathcal{X}$ is bounded, we have for any $p \geqslant 1$, using the Holder inequality:

$$\mathbb{W}_p(p_1, p_2) \leqslant \mathrm{diam}(\mathcal{X})^{(p-1)/p} \mathbb{W}_1(p_1, p_2)^{1/p}, \tag{37}$$

$$\mathbb{W}_1(p_1, p_2) \leqslant \mathrm{diam}(\mathcal{X}) d_{TV}(p_1, p_2), \tag{38}$$

$$d_{TV}(p_1, p_2) \leqslant |\mathcal{X}|^{1/2} \|p_1 - p_2\|_{L^2(\mathcal{X})}, \tag{39}$$

where $\mathrm{diam}(\mathcal{X})$ denotes the diameter of the set $\mathcal{X}$.

A.3 General PSD models

In this section, we recall the definition of a PSD model more generally as introduced by Rudi and Ciliberto [15].

Following Marteau-Ferey et al. [9], Rudi and Ciliberto [15], we consider the family of positive semi-definite (PSD) models, namely non-negative functions parametrized by a feature map $\phi : \mathcal{X} \to \mathcal{H}$ from an input space $\mathcal{X}$ to a suitable feature space $\mathcal{H}$ (a separable Hilbert space e.g. $\mathbb{R}^q$) and a linear operator $M \in \mathbb{S}_+(\mathcal{H})$, of the form

$$f(x;\ M, \phi) = \phi(x)^\top M \ \phi(x). \tag{40}$$

PSD models offer a general way to parametrize non-negative functions (since M is positive semidefinite, $f(x;\ M, \phi) \geqslant 0$ for any $x \in \mathcal{X}$) and enjoy several additional appealing properties discussed in the following. In this work. we focus on a special family of models i.e. Gaussian PSD models defined in Sec. 2 and Eq. (1). These models parametrize probability densities over $\mathcal{X} \subset \mathbb{R}^d$. It is a special case of Eq. (40) where $i)$ $\phi = \phi_\eta : \mathbb{R}^d \to \mathcal{H}_\eta$ is a feature map associated to the Gaussian kernel defined in Example 1, or by Schölkopf and Smola [18] and, $ii)$ the operator M lives in the span of $\phi(x_1), \ldots, \phi(x_n)$ for a given set of points $(x_i)_{i=1}^n$, namely there exists $A \in \mathbb{S}_+(\mathbb{R}^n)$ such that $M = \sum_{ij} A_{ij} \phi_\eta(x_i) \phi_\eta(x_j)^\top$.

Thus, given the triplet (A, X, η) characterizing the Gaussian PSD model in Eq. (1), we have

$$\sum_{1\leqslant i,j\leqslant n} A_{ij} k_\eta(x, x_i) k_\eta(x, x_j) = f(x; A, X, \eta) \qquad = f(x; M, \phi_\eta)$$
$$M = \sum_{1\leqslant i,j\leqslant n} A_{ij} \phi_\eta(x_i) \otimes \phi_\eta(x_j),$$

where $(u \otimes v)w = uv^\top w = \langle v, w\rangle\, u$.

B Properties of the Gaussian RKHS

In this section, we introduce notations and results associated to the Gaussian RKHS (see Example 1) $\mathcal{H}_\eta$ for a given $\eta \in \mathbb{R}^d_{++}$ (η will sometimes be taken in the form $\tau \mathbf{1}_d$). Recall that the Gaussian embedding is written $\phi_\eta : \mathbb{R}^d \to \mathcal{H}_\eta$ and that the Gaussian kernel is denoted with k_η.

B.1 Properties of the Gaussian kernel k_η

The following lemma has an immediate proof.

Lemma 1 (product of Gaussian kernels)**.** *Let $K \in \mathbb{N}$, let $\eta_1, ..., \eta_K \in \mathbb{R}^d_{++}$ and let $y_1, ..., y_K \in \mathbb{R}^d$. The following equality holds:*

$$\forall x \in \mathbb{R}^d,\ \prod_{k=1}^K k_{\eta_k}(x, y_k) = k_{\overline{\eta}}(x, \overline{y}) \prod_{k=1}^K k_{\eta_k}(y_k, \overline{y})$$

where $\overline{\eta} = \sum_{k=1}^K \eta_k$ and $\overline{y} = \sum_k \eta_k y_k / \overline{\eta}$

Let us now state an useful corollary.

Corollary 1. *Let $\eta \in \mathbb{R}^d_{++}$, $y_1, y_2 \in \mathbb{R}^d$. Then*

$$\forall x \in \mathbb{R}^d,\ k_\eta(x, y_1) k_\eta(x, y_2) = k_{2\eta}(x, (y_1 + y_2)/2) k_{\eta/2}(y_1, y_2). \tag{41}$$

Lemma 2 (Gaussian embedding derivative)**.** *Let $\eta \in \mathbb{R}^d_{++}$, $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{N}^d$. The derivative $\partial_\alpha \phi_\eta(x)$ is well defined in $\mathcal{H}_\eta$, and $\|\partial_\alpha \phi_\eta(x)\|_{\mathcal{H}_\eta} = 2^{|\alpha|/2} \eta^{\alpha/2}$. Moreover, if $g \in \mathcal{H}_\eta$, then $\sup_{x\in\mathbb{R}^d} |(\partial_\alpha g)(x)| \leqslant 2^{|\alpha|/2} \eta^{\alpha/2} \|g\|_{\mathcal{H}_\eta}$.*

Proof. Let $\alpha \in \mathbb{N}^d$ and let $v_\eta(z) = k_\eta(z, 0) = \exp(-z^\top \operatorname{diag}(\eta) z)$. If the function $\frac{\partial^\alpha}{\partial x^\alpha} k_\eta(x, y)$ belongs to $\mathcal{H}_\eta$, then $\partial_\alpha \phi_\eta(x)$ is in $\mathcal{H}_\eta$ and is equal to that function by the reproducing property.

First, note that

$$\forall x, y \in \mathbb{R}^d,\ \tfrac{\partial^\alpha}{\partial x^\alpha} k_\eta(x, y) = (-1)^{|\alpha|} \partial_\alpha \tau_x[v_\eta](y),$$

where $\tau_x : f \mapsto f(\cdot - x)$, commutes with the differential operator ∂_α, and satisfies the following relation wrt to the Fourier transform : $\mathcal{F}[\tau_x g](\xi) = e^{-2i\pi x\xi} \mathcal{F}[f](\xi)$. Hence, using (e) of proposition 3, we get the following fourier transform wrt y:

$$\mathcal{F}_y[\tfrac{\partial^\alpha}{\partial x^\alpha} k_\eta(x, y)](\xi) = (-2\pi i)^{|\alpha|} \xi^\alpha e^{-2i\pi\xi x} \mathcal{F}[v_\eta](\xi).$$

Hence, we have using Eq. (26):

$$\begin{aligned}\|\tfrac{\partial^\alpha}{\partial x^\alpha} k_\eta(x,\cdot)\|^2_{\mathcal{H}_\eta} &= \int_{\mathbb{R}^d} (2\pi)^{2|\alpha|}\xi^{2\alpha}\mathcal{F}[v_\eta](\xi)\,d\xi \\ &= (-1)^{|\alpha|}\int_{\mathbb{R}^d} (2i\pi)^{2|\alpha|}\xi^{2\alpha}\mathcal{F}[v_\eta](\xi)\,d\xi \\ &= (-1)^{|\alpha|}\int_{\mathbb{R}^d} \mathcal{F}[\partial_{2\alpha}v_\eta](\xi)\,d\xi = (-1)^{|\alpha|}\partial_{2\alpha}v_\eta(0),\end{aligned}$$

where the last equality comes from the inverse Fourier transform. A simple recursion then shows that $(-1)^{|\alpha|}\partial_{2\alpha}v_\eta(0) = 2^{|\alpha|}\eta^\alpha$, hence the result. The last point of the lemma is simply a consequence of the fact that $\partial_\alpha g(x) = \langle g, \partial_\alpha \phi_\eta(x)\rangle_{\mathcal{H}_\eta}$.

□

B.2 Useful Matrices and Linear Operators on the Gaussian RKHS

Recall that we denote with ϕ_η the embedding associated to the RKHS $\mathcal{H}_\eta$ of the Gaussian kernel k_η defined in Example 1. In this section, we define operators which will be useful throughout the rest of this section and which we will use in Appendixes E and F. In order to make the dependence in η appear (indeed, η will be a parameter to choose in the the next sections), we will keep it as an index for all of these operators. Recall that for any two vectors u, v in a Hilbert space $\mathcal{H}$, we can define their tensor product $u \otimes v$ which is a linear rank one operator on $\mathcal{H}$ defined by $(u \otimes v)w = \langle v, w\rangle_{\mathcal{H}} u$. For the sake of simplicity, we will often write $u \otimes v$ as $uv^\top$, so that the formula $(u \otimes v)w = uv^\top w$ is formally true.

Kernel matrices. We start off by setting the notations for kernel matrices as done by Rudi and Ciliberto [15]. Let $X \in \mathbb{R}^{n\times d}$ and $X' \in \mathbb{R}^{n'\times d}$ be two matrices corresponding to points $x_1, ..., x_n \in \mathbb{R}^d$ and $x'_1, ..., x'_{n'} \in \mathbb{R}^d$. We denote with $K_{X,X',\eta}$ the matrix in $\mathbb{R}^{n\times n'}$ such that

$$\forall 1 \leqslant i \leqslant n,\ \forall 1 \leqslant j \leqslant n',\ [K_{X,X',\eta}]_{ij} = k_\eta(x_i, x'_j). \tag{42}$$

If $X = X'$, then we just write $K_{X,\eta}$ and it is positive semi-definite, i.e. $K_{X,\eta} \in \mathbb{S}_+(\mathbb{R}^n)$.

Integration matrices. In this work, we also define, for a given hyper-rectangle $Q = \prod_{k=1}^d [a_k, b_k]$, the following integration matrix $G_{X,X',\eta,Q} \in \mathbb{R}^{n\times n'}$:

$$\begin{aligned}&\forall 1 \leqslant i \leqslant n,\ \forall 1 \leqslant j \leqslant n',\ [G_{X,X',\eta,Q}]_{ij} = \int_Q k_\eta(x - (x_i + x_j)/2)dx \\ &= \prod_{k=1}^d \sqrt{\tfrac{\pi}{4\eta_k}}\left(\operatorname{erf}(\sqrt{\eta_k}(b_k + (x_{ik} + x'_{jk})/2)) - \operatorname{erf}(\sqrt{\eta_k}(a_k + (x_{ik} + x'_{jk})/2))\right),\end{aligned} \tag{43}$$

where the erf function is defined in the notations section. Similarly, if $X = X'$, we simply write $G_{X,\eta,Q}$.

This matrix is defined in order to satisfy the following property, which is a direct application of Eq. (41): for any $X \in \mathbb{R}^{n\times d}$, any $A \in \mathbb{S}^n_+$ and $\eta \in \mathbb{R}^d_{++}$, the following holds.

$$\int_Q f(x;\ A, X, \eta)\ dx = \sum_{1 \leqslant i,j \leqslant n} [A \circ K_{X,\eta/2} \circ G_{X,2\eta,Q}]_{ij} = \mathrm{vec}(A \circ K_{X,\eta/2} \circ G_{X,2\eta,Q})^\top \mathbf{1}_{n^2} \tag{44}$$

Co-variance operator. Let $\mathcal{X} \subset \mathbb{R}^d$ be a measurable set of $\mathbb{R}^d$ with finite Lebesgue measure $|\mathcal{X}|$. Define the associated co-variance operator:

$$C_\eta \in \mathbb{S}_+(\mathcal{H}_\eta), \qquad C_\eta = \tfrac{1}{|\mathcal{X}|} \int_{\mathcal{X}} \phi_\eta(x) \otimes \phi_\eta(x)\ dx, \qquad C_{\eta,\lambda} = C_\eta + \lambda I. \tag{45}$$

Note that C_η is a trace class operator with and that $\mathrm{Tr}\,(C_\eta) = 1$ by linearity of the trace and since $\mathrm{Tr}\,(\phi_\eta(x) \otimes \phi_\eta(x)) = \|\phi_\eta(x)\|^2 = k_\eta(x,x) = 1$. Moreover, since $C_{\eta,\lambda} \succeq \lambda I$, $C_{\eta,\lambda}$ is inverible for any $\lambda > 0$.

Note that we do not make the set $\mathcal{X}$ appear in the notation of the co-variance operator (which can actually be defined with respect to any probability distribution on $\mathbb{R}^d$ and not just $\frac{\mathbf{1}_{\mathcal{X}}\ dx}{|\mathcal{X}|}$). This is because the set $\mathcal{X}$ will usually explicit in the next sections, and in particular equal to the unit hyper-cube $\mathcal{X} = (-1,1)^d$.

Sampling operators. Let $n \in \mathbb{N}$ $(x_1, .., x_n) \in (\mathbb{R}^d)^n$ be points of $\mathbb{R}^d$ which should be seen as samples from a certain distribution. We define the following sampling operators.

$$\widehat{C}_\eta \in \mathbb{S}_+(\mathcal{H}_\eta), \qquad \widehat{C}_\eta = \tfrac{1}{n} \sum_{i=1}^n \phi_\eta(x_i) \otimes \phi_\eta(x_i), \qquad \widehat{C}_{\eta,\lambda} = \widehat{C}_\eta + \lambda I \tag{46}$$

$$\widehat{S}_\eta : \mathcal{H}_\eta \to \mathbb{R}^n, \qquad \widehat{S}_\eta(g) = \tfrac{1}{\sqrt{n}} (g(x_i))_{1 \leqslant i \leqslant n} \tag{47}$$

$$\widehat{S}^*_\eta : \mathbb{R}^n \to \mathcal{H}_\eta, \qquad \widehat{S}^*_\eta(a) = \tfrac{1}{\sqrt{n}} \sum_{i=1}^n a_i \phi_\eta(x_i) \tag{48}$$

where $\widehat{S}^*_\eta$ and $\widehat{S}_\eta$ are adjoint operators. We will usually use the $\widehat{\bullet}$ notation to denote sampling operators, and imply the underlying $(x_1, ..., x_n)$. These operators will be used in later sections in order to quantify the difference between objects resulting from the sampling of distributions and the "ideal" objects (typically the difference between an empirical risk minimizer and the true expected risk minimizer). For instance, it is clear the $\widehat{C}_\eta$ is an empirical version of C_η, if the x_i are i.i.d. samples from the uniform distribution on $\mathcal{X}$.

Compression operators. Following the notations of Rudi and Rosasco [16], Rudi et al. [13], Rudi and Ciliberto [15], a compression operator of size m is an operator $\widetilde{Z}_{\eta,m} : \mathcal{H}_\eta \to \mathbb{R}^m$. We call it a *compression operator* since we use it to project every element of $\mathcal{H}_\eta$ onto the range of the adjoint operator $\widetilde{Z}^*_{\eta,m} : \mathbb{R}^m \to \mathcal{H}_\eta$. This range, which we denote with $\widetilde{\mathcal{H}}_{\eta,m} \subset \mathcal{H}_\eta$, is a subset of dimension at most m. We also denote with $\widetilde{P}_{\eta,m} : \mathcal{H}_\eta \to \mathcal{H}_\eta$ the orthogonal projection onto $\widetilde{\mathcal{H}}_{\eta,m}$, which can also be written $\widetilde{P}_{\eta,m} = \widetilde{Z}^*_{\eta,m} (\widetilde{Z}_{\eta,m} \widetilde{Z}^*_{\eta,m})^\dagger \widetilde{Z}_{\eta,m}$, where $\dagger$ denotes the Moore-Penrose pseudo-inverse.

In this work, we will always use the notation $\widetilde{\bullet}_m$ to denote a compression operator, and the index m to make the size of the compression explicit.

In this work, we take a specific form of compression operator as in appendix C of Rudi and Ciliberto [15]. Indeed, let $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ be a data point matrix representing vectors $\tilde{x}_1, ..., \tilde{x}_m \in \mathbb{R}^d$. The compression operator associated to $\widetilde{X}_m$ is the following :

$$\widetilde{Z}_{\eta,m} : \mathcal{H}_\eta \to \mathbb{R}^m, \qquad \widetilde{Z}_{\eta,m}(g) = (g(\tilde{x}_j))_{1\leqslant j\leqslant m} = (g^\top \phi_\eta(\tilde{x}_j))_{1\leqslant j\leqslant m}. \tag{49}$$

Note that $\widetilde{Z}_{\eta,m}\widetilde{Z}^*_{\eta,m} = K_{\widetilde{X}_m,\eta}$ and hence the projection operator can be written $\widetilde{P}_{\eta,m} = \widetilde{Z}^*_{\eta,m}K^\dagger_{\widetilde{X}_m,\eta}\widetilde{Z}_{\eta,m}$ and that it is simply the projection onto $\mathrm{span}\{\phi_\eta(\tilde{x}_i)\}_{1\leqslant i\leqslant m}$. This compression is also chosen to satisfy the two following properties :

- if $h \in \mathcal{H}_\eta$, then $\widetilde{P}_{\eta,m}h$ represents a function of the form $g(\bullet;\ a, \widetilde{X}_m, \eta)$ where $a = K^\dagger_{\widetilde{X}_m,\eta}\widetilde{Z}_{\eta,m}h$ (see Eq. (2) for the definition of the Gaussian linear model $g(x;\ a, \widetilde{X}_m, \eta)$);
- if $M \in \mathbb{S}_+(\mathcal{H}_\eta)$, then for any $x \in \mathbb{R}^d$, it holds

$$f(x;\ \widetilde{P}_{\eta,m}M\widetilde{P}_{\eta,m}, \phi_\eta) = f(x;\ A, \widetilde{X}_m, \eta), \qquad A = K^\dagger_{\widetilde{X}_m,\eta}\widetilde{Z}_{\eta,m}M\widetilde{Z}^*_{\eta,m}K^\dagger_{\widetilde{X}_m,\eta}, \tag{50}$$

meaning that compressed linear (resp. PSD) models can be compressed as a sum of m (resp. m^2) Gaussian kernel functions. We quantify the effect of this compression in Lemma 3 and Theorem 5.

B.3 Approximation properties of the Gaussian kernel

This section aims in quantifying the approximation power of the Gaussian RKHS. We start in proposition 7 by quantifying the approximation power of the Gaussian RKHS by finding an ε approximation of a regular function with controlled norm. We then quantify the "size" of a compression for the Gaussian RKHS in Lemma 3, which essentially bounds the possible variations of a function in $\mathcal{H}_\eta$ if it is equal to zero on the compression points $\widetilde{X}_m$.

Approximation of a Sobolev function. This paragraph remolds results in the proof of Theorem D.4 of Rudi and Ciliberto [15] whose goal is to approximate any function $g \in W^\beta_2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ by a function in $\mathcal{H}_\eta$.

Proposition 7 (Approximation of $W^\beta_2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ in $\mathcal{H}_\eta$)**.** *Let g be a function in $W^\beta_2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ and $\eta \in \mathbb{R}^d_{++}$. Denote with $|\eta|$ the product $|\eta| := \prod_{i=1}^d \eta_i$ and $\eta_0 = \min_{1\leqslant i\leqslant d}\eta_i$. For any $\varepsilon \in (0, 1]$, there exists $\theta \in \mathcal{H}_\eta$ such that*

$$\begin{cases} \|\theta - g\|_{L^2(\mathbb{R}^d)} \leqslant \varepsilon \|g\|_{W^\beta_2(\mathbb{R}^d)} \\ \|\theta - g\|_{L^\infty(\mathbb{R}^d)} \leqslant C_1\, \varepsilon^{1-\nu} \|g\|_\bullet \end{cases}, \quad \|\theta\|_{\mathcal{H}_\eta} \leqslant C_2 \|g\|_{W^\beta_2(\mathbb{R}^d)} |\eta|^{1/4} \left(1 + \varepsilon \exp\left(\tfrac{50}{\eta_0 \varepsilon^{2/\beta}}\right)\right), \tag{51}$$

where $\|g\|_\bullet = \|g\|_{L^\infty(\mathbb{R}^d)}$ if $\beta \leqslant d/2$ and $\|g\|_\bullet = \|g\|_{W^\beta_2(\mathbb{R}^d)}$ if $\beta > d/2$, $\nu = \min(1, d/(2\beta))$ and C_1, C_2 are constants which depend only on d, β.

Proof. Recalling the notations from the proof of Theorem D.4. of Rudi and Ciliberto [15], let $g_t := t^{-d}g_1(x/t)$ where g_1 is defined as g in equation (D.2) of Rudi and Ciliberto [15]. The following result hold.

- By step 1 of the proof of Theorem D.4, $\|g - g \star g_t\|_{L^2(\mathbb{R}^d)} \leqslant (2t)^\beta \|g\|_{W^\beta_2(\mathbb{R}^d)}$.

- By step 2 and the beginning of step 3 of the proof of Theorem D.4,

$$\|g \star g_t\|_{\mathcal{H}_\eta} \leqslant 2^\beta \pi^{-d/4} |\eta|^{1/4} (1 + (t/3)^\beta \exp(\tfrac{50}{\eta_0 t^2})) \|g\|_{W_2^\beta(\mathbb{R}^d)}.$$

- As in step 5 of the proof of Theorem D.4 and in particular the Young convolution inequality combined with the fact that $\|g_1\|_{L^1(\mathbb{R}^d)}$ is finite, $\|g \star g_t\|_{L^\infty(\mathbb{R}^d)} \leqslant \|g_1\|_{L^1(\mathbb{R}^d)} \|g\|_{L^\infty(\mathbb{R}^d)}$ which in turn implies $\|g - g \star g_t\| \leqslant (1 + \|g_1\|_{L^1(\mathbb{R}^d)}) \|g\|_{L^\infty(\mathbb{R}^d)}$.

Replacing t by $\varepsilon^{1/\beta}/2$, we get all the bounds except the bound for the L^∞ norm in the case where $\beta > d/2$. In that case, we proceed in the following way. Recycling results and notations from the proof of Theorem D.4 of Rudi and Ciliberto [15], denoting with $\mathcal{F}$ the Fourier transform defined in proposition 3, it holds

$$\begin{aligned}
\|f - f \star g_t\|_{L^\infty(\mathbb{R}^d)} &\leqslant \|\mathcal{F}(f - f \star g_t)\|_{L^1(\mathbb{R}^d)} \text{ proposition 3} \\
&= \|\mathcal{F}(f)(1 - \mathcal{F}(g_t))\|_{L^1(\mathbb{R}^d)} \\
&\leqslant \|(1 + \|\omega\|^2)^{\beta/2} \mathcal{F}(f)\|_{L^2(\mathbb{R}^d)} \, \|(1 + \|\omega\|^2)^{-\beta/2} \, \mathcal{F}(1 - g_t)\|_{L^2(\mathbb{R}^d)} \\
&\leqslant 2^\beta \left(\int_{\|\omega\| > 1/t} (1 + \|\omega\|^2)^{-\beta} \, d\omega \right)^{1/2} \|f\|_{W_2^\beta(\mathbb{R}^d)} \text{ Eq. (29)} \\
&= 2^\beta \left(S_d \int_{r > 1/t} r^{d-1} \, (1 + r^2)^{-\beta} \, dr \right)^{1/2} \|f\|_{W_2^\beta(\mathbb{R}^d)} \text{ (spherical coord.)} \\
&\leqslant 5^{\beta/2} S_d^{1/2} \left(\int_{r > 1/t} r^{d-1-2\beta} \, dr \right)^{1/2} \|f\|_{W_2^\beta(\mathbb{R}^d)} \; (t < 1/2) \\
&= 5^{\beta/2} \tfrac{1}{\sqrt{2\beta - d}} S_d^{1/2} t^{\beta - d/2} \|f\|_{W_2^\beta(\mathbb{R}^d)} \\
&= 5^{\beta/2} 2^{d/2 - \beta} S_d^{1/2} \tfrac{1}{\sqrt{2\beta - d}} \varepsilon^{1 - d/(2\beta)} \|f\|_{W_2^\beta(\mathbb{R}^d)},
\end{aligned}$$

where S_d is the surface area of the $d-1$ dimensional hyper-sphere. □

A bound on the performance of compression when using uniform samples from $\mathcal{X} = (-1, 1)^d$. In this paragraph, we study the effect of performing compression with a compression operator of the form $\widetilde{Z}_{\eta,m}$ (see Eq. (49)) where the associated $\widetilde{X}_m$ are i.i.d. samples from the uniform measure on the unit hyper-cube $\mathcal{X} = (-1, 1)^d$.

Lemma 3. *Let $m \in \mathbb{N}$, $\delta \in (0, 1]$, $\tau \geqslant 1$ and $\rho \in (0, 1]$. Let $\eta = \tau \mathbf{1}_d \in \mathbb{R}^d_{++}$. Let $\widetilde{X}_m \in \mathbb{R}^{m \times d}$ be a data matrix corresponding to vectors $\tilde{x}_1, ..., \tilde{x}_m$ which are sampled independently and uniformly from $\mathcal{X} = (-1, 1)^d$ and let $\widetilde{P}_{\eta,m}$ be the associated projection operator in $\mathcal{H}_\eta$. With probability at least $1 - \delta$, if $m \geqslant C_1 \tau^{d/2} (\log \frac{C_2}{\rho})^d \left(\log \frac{C_3}{\delta} + \log \tau + \log \log \frac{C_2}{\rho} \right)$, then it holds :*

$$\sup_{x \in \mathcal{X}} \|(I - \widetilde{P}_{\eta,m}) \phi_\eta(x)\| \leqslant \rho, \tag{52}$$

where C_1, C_2, C_3 are constants which depend only on the dimension d and not on τ, m, δ, ρ.

Proof. Let h denote the fill distance with respect to $\widetilde{X}_m$, i.e.

$$h = \max_{x \in [-1,1]^d} \min_{1 \leqslant j \leqslant m} \|x - \tilde{x}_j\| \tag{53}$$

Using Lemma 12 p.19 of Vacher et al. [23], we there exists two constants C_1, C_2 depending only on d such that $h \leqslant \left(C_1 m^{-1}(\log(C_2 m/\delta)\right)^{1/d}$.

Applying Theorem C.3 from Rudi and Ciliberto [15] in the case where $\mathcal{X} = (-1,1)^d$, $\eta = \tau \mathbf{1}_d$, there exists constants C_3, C_4, C_5 depending only on the dimension d such that when $h \leqslant \tau^{-1/2} C_3^{-1}$, the following holds :

$$\sup_{x \in \mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| \leqslant C_4 e^{-\frac{C_5}{\tau^{1/2}h} \log \frac{C_5}{\tau^{1/2}h}} \tag{54}$$

Now note that taking $C_6 = \max(C_3^{-1}, eC_5)$ and $C_7 = \max(e, C_4)$, as soon as $h \leqslant C_6 \tau^{-1/2} / \log \frac{C_7}{\rho}$, it holds a) $h \leqslant \tau^{-1/2} C_3^{-1}$, b) $\frac{C_5}{\tau^{1/2}h} \geqslant e$ and thus $\log \frac{C_5}{\tau^{1/2}h} \geqslant 1$, and hence c) $\sup_{x \in \mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| \leqslant \rho$ using Eq. (54). Using the bound on h, this is satisfied as soon as

$$m \geqslant C_8 \tau^{d/2} \left(\log \frac{C_7}{\rho}\right)^d \log(C_2 m/\delta),$$

where $C_8 = \max(C_1/C_6^d, e)$. Using the fact that $C_2, C_8 \geqslant e$, and using the reasoning in the proof of Theorem C.5 of Rudi and Ciliberto [15], in equation (C.44), a sufficient condition is the following :

$$m \geqslant 2C_8 \tau^{d/2} \left(\log \frac{C_7}{\rho}\right)^d \left(\log(2C_2 C_8/\delta) + \frac{d}{2} \log \tau + d \log \log \frac{C_7}{\rho}\right). \tag{55}$$

The result in the theorem is obtained by taking $C_1 \leftarrow 2C_8 d,\ C_2 \leftarrow C_7,\ C_3 \leftarrow 2C_2 C_8$. □

C Properties of Gaussian PSD models

In this section, we detail some of the properties specific to Gaussian PSD models.

C.1 Bounds on the support and the derivatives

In this section, we present results which can be used to bound the tail and derivatives of a Gaussian PSD model. These bounds can be used both for theoretical purposes (see Appendixes E and F) and to perform adaptive bounds in an algorithm (see Sec. 3)

Lemma 4 (tail bound)**.** *Let $\delta = (\delta_k) \in \mathbb{R}^d$, $\eta \in \mathbb{R}^d_{++}$, $X \in \mathbb{R}^{n \times d}$ and $A \in \mathbb{S}_+(\mathbb{R}^n)$. Let $f(x; A, X, \eta)$ be the associated PSD model. Define $\overline{x}$, $\underline{x}$:*

$$\forall 1 \leqslant k \leqslant d,\ \overline{x}_k = \max_{1 \leqslant i \leqslant n} X_{ik},\ \underline{x}_k = \min_{1 \leqslant i \leqslant n} X_{ik}.$$

Let $Q_\delta = Q(\underline{x} - \delta, \overline{x} + \delta)$. Then the following bound holds:

$$\int_{\mathbb{R}^d \setminus Q_\delta} |f(x; A, X, \eta)| dx \leqslant \left(2\pi^{d/2} \det(\operatorname{diag}(2\eta))^{-1/2} \sum_{k=1}^d e^{-2\eta_k \delta_k^2}\right) \sum_{i,j} [A \circ K_{X,\eta/2}]_{ij} \tag{56}$$

Proof. Start by recalling the following simple Chernoff bound:

$$\forall x > 0,\ \int_x^{+\infty} e^{-t^2}\, dt \leqslant \sqrt{\pi} e^{-x^2} \tag{57}$$

Indeed, take $\lambda > 0$. Since $e^{-2\lambda x}\, e^{2\lambda t} \leqslant \mathbf{1}_{t>x}$, it holds

$$\int_x^{+\infty} e^{-t^2}\, dt \leqslant e^{-2\lambda x} e^{\lambda^2} \int_{-\infty}^{+\infty} e^{-(t-\lambda)^2}\, dt \leqslant \sqrt{\pi} e^{-x^2} e^{(\lambda - x)^2}.$$

Hence, taking $\lambda = x$, we get the bound. Then we perform the following bound.

$$\begin{aligned}
\int_{\mathbb{R}^d \setminus Q(-\delta,\delta)} k_\eta(x,0)dx &= \frac{1}{\prod_{k=1}^d \eta_k^{1/2}} \int_{\mathbb{R}^d \setminus Q(-\delta\sqrt{\eta},\delta\sqrt{\eta})} k_1(x,0)dx \\
&\leqslant \frac{1}{\prod_{k=1}^d \eta_k^{1/2}} \sum_{k=1}^d \left(\pi^{(d-1)/2} 2 \int_{\delta_k\sqrt{\eta_k}}^{\infty} e^{-t^2}\, dt \right) \\
&\leqslant 2\pi^{d/2} \det(\mathrm{diag}(\eta))^{-1/2} \sum_{k=1}^d e^{-\delta_k^2 \eta_k},
\end{aligned}$$

where we go from the first to the second line by noting that

$$\mathbb{R}^d \setminus Q(-\delta,\delta) \subset \cup_{k=1}^d \mathbb{R} \times .... \times \mathbb{R} \setminus [-\delta_k, \delta_k] \times ... \times \mathbb{R},$$

and the last inequality comes from a Eq. (57).

The result immediately follows from Eq. (41) as well as the fact that Q_δ contains $(x_i + x_j)/2 + Q(-\delta,\delta)$ for all $1 \leqslant i,j \leqslant n$.

□

Lemma 5 (derivative bound for general PSD model)**.** *Let $\eta \in \mathbb{R}^d_{++}$, $M \in \mathbb{S}_+(\mathcal{H}_\eta)$ $X \in \mathbb{R}^{n\times d}$ and $A \in \mathbb{S}^n_+$. The following bounds hold :*

$$\sup_{x\in\mathbb{R}^d} |\partial_\alpha f(x;\ M, \phi_\eta)| \leqslant 2^{3|\alpha|/2}\, \eta^{\alpha/2}\, \|M\| \tag{58}$$

$$\sup_{x\in\mathbb{R}^d} |\partial_\alpha f(x;\ A, X, \eta)| \leqslant 2^{3|\alpha|/2}\, \eta^{\alpha/2}\, \|K_{X,\eta}^{1/2} A K_{X,\eta}^{1/2}\| \tag{59}$$

Proof. By derivation of a bi-linear form, we get

$$\partial_\alpha f(x;\ M, \phi_\eta) = \sum_{\beta\leqslant\alpha} \binom{\alpha}{\beta} \langle \partial_\beta \phi_\eta(x), M \partial_{\alpha-\beta}\phi_\eta(x) \rangle_{\mathcal{H}_\eta}$$

Hence, using Lemma 2, we get, for any $x \in \mathbb{R}^d$,

$$|\partial_\alpha f(x;\ M, \phi_\eta)| \leqslant \|M\| \sum_{\beta\leqslant\alpha} \binom{\alpha}{\beta} 2^{|\beta|/2} \eta^{\beta/2} 2^{|\alpha-\beta|/2} \eta^{(\alpha-\beta)/2} = 2^{3|\alpha|/2} \eta^{\alpha/2} \|M\|. \tag{60}$$

In particular, since $f(x;\ A, X, \eta) = f(x;\ M_A, \phi_\eta)$ with $M_A = Z^* A Z$ for $Z : h \in \mathcal{H}_\eta \mapsto h(x_i)_{1\leqslant i\leqslant n}$, and since $ZZ^* = K_{X,\eta}$, it holds

$$\|M_A\| = \|Z^* A Z\| = \|A^{1/2} Z Z^* A^2\| = \|A^{1/2} K_{X,\eta} A^{1/2}\| = \|K_{X,\eta}^{1/2} A K_{X,\eta}^{1/2}\|,$$

and hence the second equation of the lemma. □

C.2 Compression as a Gaussian PSD model

In this section, we restate Theorem C.4 of Rudi and Ciliberto [15] on the compression of a PSD model of the form $f(x;\ M, \phi_\eta)$ into a Gaussian PSD model.

Let $\eta \in \mathbb{R}^d_{++}$, $M \in \mathbb{S}_+(\mathcal{H}_\eta)$. Given a matrix $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ representing vectors $\tilde{x}_1, ..., \tilde{x}_m \in \mathbb{R}^d$, and the associated projection operator $\widetilde{P}_{\eta,m}$ (for more details, see Appendix B.2), one can compress the PSD model $f(\bullet;\ M, \phi_\eta)$ into $f(\bullet;\ \widetilde{P}_{\eta,m} M \widetilde{P}_{\eta,m})$ which is also a Gaussian PSD model of the form $f(\bullet;\ A, \widetilde{X}_m, \eta)$ (A is defined in Eq. (50)). The quality of the compression is given by the following theorem.

Theorem 5 (Theorem C.4 of Rudi and Ciliberto [15])**.** *Using the previous notations, the compressed model associated to $\widetilde{P}_{\eta,m} M \widetilde{P}_{\eta,m}$ of M onto $\widetilde{X}_m$ has a distance to the original PSD model associated to M bounded, for any $x \in \mathcal{X}$, by*

$$|f(x;\ M, \phi_\eta) - f(x;\ \widetilde{P}_{\eta,m} M \widetilde{P}_{\eta,m}, \phi_\eta)| \leqslant \sqrt{f(x;\ M, \phi_\eta)}\ \|M\|^{1/2} \sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| + \|M\| \sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\|^2. \tag{61}$$

We therefore see that the quality of the compression depends mainly on the quantity

$$\sup_{x\in\mathcal{X}} \|(I - \widetilde{P})\phi_\eta(x)\|,$$

which can be bounded using Eq. (52) in Lemma 3.

C.3 Approximation properties of Gaussian PSD model

Define, for any measurable $\Omega \subset \mathbb{R}^d$, and any $f : \Omega \to \mathbb{R}$, the following function (set to $+\infty$ if the set is empty).

$$\|f\|_{\mathsf{sos},\Omega,\beta} = \inf \left\{ \sum_{i=1}^{Q} \max(\|f_j\|_{L^\infty(\Omega)}, \|f_j\|_{W_2^\beta(\Omega)})^2 \mid f = \sum_{j=1}^{Q} f_j^2,\ Q \in [0, +\infty] \right\} \tag{62}$$

Here, we recall Theorem D.4 of Rudi and Ciliberto [15], refined in a small way to have more control over the dependence in the f_j.

Theorem 6 (Theorem D.4 of Rudi and Ciliberto [15])**.** *Let $\tau \geqslant 1$ and $\varepsilon \in (0, 1]$ and f such that $\|f\|_{\mathsf{sos},\mathbb{R}^d,\beta} < \infty$. Let $\eta = \tau \mathbf{1}_d$. There exists $M_{\tau,\epsilon} \in \mathbb{S}_+(\mathcal{H}_\eta)$ such that $f_{\tau,\epsilon} := f(\bullet;\ M_{\tau,\epsilon}, \phi_\eta)$ is ε close to p in L^2 norm and has controlled trace norm:*

$$\begin{aligned} &\|f_{\tau,\epsilon} - f\|_{L^2(\mathbb{R}^d)} \leqslant C_1\ \|f\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ \varepsilon, \\ &\mathrm{Tr}\,(M_{\tau,\epsilon}) \leqslant C_2\ \|f\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ \tau^{d/2}(1 + \varepsilon^2 \exp(C_3\, \varepsilon^{-2/\beta}/\tau)), \end{aligned} \tag{63}$$

where the constants C_1, C_2, C_3 depend only on β, d.

Proof. Let $\delta > 0$ and take $Q_\delta \in [0, +\infty]$ as well as $f_{\delta,j}$ such that $f = \sum_{j=1}^{Q_\delta} f_{\delta,j}^2$ point-wise and

$$\sum_{i=1}^{Q} \|f_{\delta,j}\|_{W_2^\beta(\mathbb{R}^d)} \max(\|f_{\delta,j}\|_{L^\infty(\mathbb{R}^d)}, \|f_{\delta,j}\|_{W_2^\beta(\mathbb{R}^d)}) \leqslant \|f\|_{\mathsf{sos},\beta}\,.$$

Now using exactly the same reasoning than in the proof of Theorem D.4 of Rudi and Ciliberto [15] but setting simply $t = \varepsilon^{1/\beta}$, it holds the existence of $M_{\delta,\tau,\varepsilon}$ and C_1, C_2, C_3 depending only on β, d such that

$$\|f_{\delta,\tau,\varepsilon} - f\|_{L^2(\mathbb{R}^d)} \leqslant C_1 \left(\|f\|_{\mathsf{sos},\mathbb{R}^d,\beta} + \delta\right) \varepsilon,$$
$$\mathrm{Tr}\left(M_{\delta,\tau,\varepsilon}\right) \leqslant C_2 \left(\|f\|_{\mathsf{sos},\mathbb{R}^d,\beta} + \delta\right) \tau^{d/2} (1 + \varepsilon^2 \exp(C_3\, \varepsilon^{-2/\beta}/\tau)).$$

Note that in the proof, $M_{\delta,\tau,\varepsilon}$ is well defined since its trace norm is bounded (normal convergence). Now if $\|f\|_{\mathsf{sos},\mathbb{R}^d,\beta} = 0$, then $f = 0$ and there is nothing to prove. If not, then taking $\delta = \|f\|_{\mathsf{sos},\mathbb{R}^d,\beta}$, the theorem holds. □

D The sampling algorithm

In this section, we formally prove that algorithm 1 converges, as in Theorem 1, as well as the different results of Sec. 3. We start by introducing some notations around dyadic decomposition of hyper-rectangles. We then introduce a well founded order relation, which we will then use to both construct the random variables we study, justify the convergence of the algorithm and prove its correctness.

Recall we are given a density (up to a scaling factor) $f(x)$ and that we denote with $I(Q)$ the quantity $\int_Q f(x)dx$ on any hyper-rectangle Q.

D.1 Dyadic decompositions and convergence of algorithm 1

Dyadic sub-rectangles Let $Q = \prod_{k=1}^d [a_k, b_k[$ be a hyper-rectangle where $a \leqslant b$ and let $\delta = b - a$. Let $q \in \mathbb{N}^d$. We define $\mathcal{D}_{Q,q}$ to be the set of dyadic sub-rectangles of Q whose k-th size is cut in half q_k times, i.e.

$$\mathcal{D}_{Q,q} = \left\{ \prod_{k=1}^d [a_k + \delta_k \tfrac{s}{2^{q_k}}, a_k + \delta_k \tfrac{s+1}{2^{q_k}}[\;:\; s \in \prod_{k=1}^d [\![0, 2^{q_k} - 1]\!] \right\}.$$

We denote with $\mathcal{D}_Q$ the set of dyadic sub-rectangles of Q, i.e. the union $\bigcup_{q \in \mathbb{N}^d} \mathcal{D}_{Q,q}$.

Moreover, if $q_k^\rho = \max(0, \lceil \log_2 \frac{\delta_k}{\rho} \rceil)$, we also define $\mathcal{D}_{Q,\varepsilon} := \mathcal{D}_{Q,q^\rho}$ to be the set of dyadic sub-rectangles whose size is just below ρ.

Well founded order relation on hyper-rectangles For all $\rho > 0$, we define the following strict order relation. We say that $Q \prec_\rho Q'$ if the following conditions hold :

1. $Q \in \mathcal{D}_{Q'}$;
2. There exists $k \in [\![1, d]\!]$ such that $\delta'_k > \rho$ and $\delta_k < \delta'_k$.

This relation is obviously transitive. Moreover, if $s(Q) := \sum_{k=1}^d \delta_k(Q)$, it is easy to show that $Q \prec_\rho Q'$ implies $s(Q) \leqslant s(Q') - \rho/2$. Since $s \geqslant 0$, this in turn shows that any strictly decreasing sequence for $\prec_\rho$ is finite, and that $Q \prec_\rho Q'$ and $Q' \prec_\rho Q$ are incompatible.

We are now ready to define the random variable $\boldsymbol{Y}_{\rho,Q,n}$ by structural induction on Q for any $n \in \mathbb{N}$. Recall that for $\Omega \subset \mathbb{R}^d$, we denote with $\mathcal{U}_\Omega$ the uniform law on Ω.

Definition of the random variable $\boldsymbol{Y}_{\rho,Q,n}$ and relation to the algorithm We now define a random variable from whose distribution we sample when SamplerRec in 1 is applied.

- If $\delta(Q) \leqslant \rho$, then for any $n \in \mathbb{N}$, $\boldsymbol{Y}_{\rho,Q,n} \sim \mathcal{U}_Q^{\otimes n}$
- Else, let $n \in \mathbb{N}$ and $k_Q = \min \operatorname{argmax}_{1 \leqslant k \leqslant d} \delta_k(Q)$ be the smallest index amongst the largest sides of Q. Define Q_1 and Q_2 to be the two hyper-rectangles obtained by cutting Q in half along the direction k_Q. Since $\delta_{k_Q} > \rho$ and Q_1, Q_2 are dyadic sub-rectangles of Q, we have $Q_1, Q_2 \prec_\rho Q$.

  By structural induction, we give ourselves a probability space on which we take we take the following random variables to be independent : $\boldsymbol{Y}_{1,m} \sim \boldsymbol{Y}_{\rho,Q_1,m}$, $\boldsymbol{Y}_{2,m} \sim \boldsymbol{Y}_{\rho,Q_2,m}$ for $0 \leqslant m \leqslant n$ and $M \sim \mathcal{B}(n, I(Q_1)/I(Q))$ and define

$$\boldsymbol{Y}_{\rho,Q,n} = (\boldsymbol{Y}_{\rho,Q_1,M}, \boldsymbol{Y}_{\rho,Q_2,n-M}) := \sum_{m=0}^{n} \mathbf{1}_{M=m} (\boldsymbol{Y}_{1,m}, \boldsymbol{Y}_{2,n-m}). \tag{64}$$

Lemma 6 (Termination of the algorithm and first result)**.** *For any inputs $\rho > 0$, hyper-rectangle Q and $n \in \mathbb{N}$,* SamplerRec *in algorithm 1 terminates and returns a sample $(y_1, ..., y_n)$ from $\boldsymbol{Y}_{\rho,Q,n}$.*

Proof. This is a simple application of structural induction on the well-founded order $\prec_\rho$ for the termination and then again for the fact that a sample $(y_1, ..., y_n)$ from $\boldsymbol{Y}_{\rho,Q,n}$, using the definition of $\boldsymbol{Y}$ above. □

D.2 Proof of Theorem 1

In this section, we prove Theorem 1. To do so, we define a random variable $X_{\rho,Q}$, compute its density with respect to the Lebesgue measure on the hyper-rectangle Q (and show it is our target density), and show that $\boldsymbol{Y} = X$ up to some random shuffling.

Definition of the variable $X_{\rho,Q}$ Recall the definition of $\mathcal{D}_{Q,\rho}$ from Appendix D.1. We define a random variable $R_{\rho,Q}$ on $\mathcal{D}_{Q,\rho}$ whose law is defined $P(R_{\rho,Q} = r) = I(r)/I(Q)$. Recall that for any $r \subset \mathbb{R}^d$, we denote with $\mathcal{U}_r$ the uniform law on r. We give ourselves a measure space on which there exists a family of random variables $U_r \sim \mathcal{U}_r$ for $r \in \mathcal{D}_{Q,\rho}$ and $R \sim R_{\rho,Q}$ which are all independent and define

$$X_{\rho,Q} = U_R := \sum_{r \in \mathcal{D}_{Q,\rho}} \mathbf{1}_{R=r} U_r \tag{65}$$

Lemma 7 (density of $X_{\rho,Q}$)**.** *The density of $X_{\rho,Q}$ with respect to the Lebesgue measure is given by Eq. (7), i.e.*

$$\forall x \in Q,\ p_{X_{\rho,Q}}(x) = \sum_{r \in \mathcal{D}_{Q,\rho}} \frac{I(r)}{I(Q)} \frac{\mathbf{1}_r(x)}{|r|}. \tag{66}$$

Proof. For any measurable function f, it holds

$$\begin{aligned}\mathbb{E}[f(X_{\rho,Q})] &= \sum_{r\in\mathcal{D}_{Q,\rho}} \mathbb{E}[\mathbf{1}_{R=r} f(U_r)] \\ &= \sum_{r\in\mathcal{D}_{Q,\rho}} P(R=r)\mathbb{E}[f(U_r)] \\ &= \sum_{r\in\mathcal{D}_{Q,\rho}} \frac{I(r)}{I(Q)} \int_{\mathbb{R}^d} f(x) \frac{\mathbf{1}_r(x)}{|r|} dx \\ &= \int_{\mathbb{R}^d} f(x) \left(\sum_{r\in\mathcal{D}_{Q,\rho}} \frac{I(r)}{I(Q)} \frac{\mathbf{1}_r(x)}{|r|} \right) dx\end{aligned}$$

□

Action of a permutation and decomposition Let $n \in \mathbb{N}$. For any permutation $\tau \in \mathfrak{S}_n$ and vector $v \in \mathbb{R}^n$, denote with $\tau \star v$ the permuted vector $(v_{\tau^{-1}(i)})_{1\leqslant i\leqslant n}$.

We now define a decomposition of a permutation of n variables as i) a permutation of the first m variables and a permutation of the last $n-m$ variables ii) followed by a rearrangement of these variables.

Given $I \subset [\![1,n]\!]$ of size m, define τ_I as the unique permutation satisfying $I = \{\tau_I(1), ..., \tau_I(m)\}$, $I^c = \{\tau_I(m+1), ..., \tau_I(n)\}$ and $\tau_I(1) < ... < \tau_I(m)$ and $\tau_I(m+1) < ... < \tau_I(n)$. For any $m \in [\![0,n]\!]$, if $\mathcal{P}_m(n)$ denotes the set of subsets of $\{1,...,n\}$ of size m, the map from $\mathcal{P}_m(n) \times \mathfrak{S}_m \times \mathfrak{S}_{n-m}$ to $\mathfrak{S}_n$ defined as

$$(I, \sigma_m, \sigma_{n-m} \mapsto \left(i \mapsto \begin{cases} \tau_I(\sigma_m(i)) & \text{if } i \leqslant m \\ \tau_I(m + \sigma_{n-m}(i-m)) & \text{otherwise} \end{cases} \right) \tag{67}$$

is a bijection.

Lemma 8. *Let $\rho > 0$. Let $n \in \mathbb{N}$, Q be a hyper-rectangle of $\mathbb{R}^d$. Let σ be a random permutation independent of $\boldsymbol{Y}_{\rho,Q,n}$. Then $(\boldsymbol{Y}^{\sigma(i)}_{\rho,Q,n})_{1\leqslant i\leqslant n} \sim X^{\otimes n}_{\rho,Q}$.*

Proof. Once again, we prove this by structural induction. Fix $\rho > 0$. We will prove the following property by structural induction on the set of hyper-rectangles Q equipped with the strict order relation $\prec_\rho$:

For any $n \in \mathbb{N}$, if σ is a random permutation (i.e. distributed uniformly amongst all permutations in $\mathfrak{S}_n$), $\boldsymbol{Y}_{Q,n} \sim \boldsymbol{Y}_{\rho,Q,n}$ and both random variables are independent, then $(\boldsymbol{Y}^{\sigma(i)}_{Q,n})_{1\leqslant i\leqslant n} \sim X^{\otimes n}_{Q,\rho}$.

1) If $\delta(Q) \leqslant \rho$.

On the one hand, by definition of $\boldsymbol{Y}_{\rho,Q,n}$, it holds that for any $n \in \mathbb{N}$, $\boldsymbol{Y}_{\rho,Q,n} \sim \mathcal{U}_Q^{\otimes n}$ and hence $\boldsymbol{Y}_{Q,n} \sim \mathcal{U}_Q^{\otimes n}$. By invariance of the product measure by permutation, it also holds that $(\boldsymbol{Y}^{\sigma(i)}_{Q,n})_{1\leqslant i\leqslant n} \sim \mathcal{U}_Q^{\otimes n}$.

On the other hand, since $\delta(Q) \leqslant \rho$, it is easy to see that $q^\rho = 0$ and hence $\mathcal{D}_{Q,\rho} = \{Q\}$. Hence, by definition of $X_{\rho,Q}$ in Eq. (65), R is deterministic and hence $X_{\rho,Q} = U_Q \sim \mathcal{U}_Q$.

Putting things together, this yields $(\boldsymbol{Y}^{\sigma(i)}_{Q,n})_{1\leqslant i\leqslant n} \sim X^{\otimes n}_{\rho,Q}$.

2) Assume $\delta(Q) > \rho$ and take $n \in \mathbb{N}$. By definition of $\boldsymbol{Y}_{\rho,Q,n}$ in Eq. (64), and since our property only concerns a convergence in law, we can assume that $\boldsymbol{Y}_{Q,n}$ is of the form

$$\boldsymbol{Y}_{Q,n} = \sum_{m=0}^{n} \mathbf{1}_{M=m}(\boldsymbol{Y}_{Q_1,m}, \boldsymbol{Y}_{Q_2,n-m}),$$

where $\boldsymbol{Y}_{Q_1,m}, \boldsymbol{Y}_{Q_2,m}$ and M are independent and independent of σ, $\boldsymbol{Y}_{Q_1,m} \sim \boldsymbol{Y}_{\rho,Q_1,m}$, $\boldsymbol{Y}_{Q_2,m} \sim \boldsymbol{Y}_{\rho,Q_2,m}$ for $0 \leqslant m \leqslant n$ and $M \sim \mathcal{B}(n, I(Q_1)/I(Q))$, and Q_1, Q_2 are defined just before Eq. (64). It is easy to see that since $Q_1 \sqcup Q_2 = Q$ and $Q_1, Q_2 \prec_\rho Q$, it holds $\mathcal{D}_{Q,\rho} = \mathcal{D}_{Q_1,\rho} \sqcup \mathcal{D}_{Q_2,\rho}$ where $\sqcup$ symbolises a disjoint union.

Fix a measurable function f. Using the independence of M from the other variables and the fact that it is discrete, it holds

$$\mathbb{E}[f(\sigma \star \boldsymbol{Y}_{Q,n})] = \sum_{m=0}^{n} P(M=m)\mathbb{E}[f(\sigma \star (\boldsymbol{Y}_{Q_1,m}, \boldsymbol{Y}_{Q_2,n-m}))].$$

Now note that using our bijection Eq. (67), it holds

$$\begin{aligned}
&\mathbb{E}_{\sigma,\mathbf{Y}_{Q_1,m},\mathbf{Y}_{Q_2,m}}\left[f(\sigma \star (\boldsymbol{Y}_{Q_1,m}, \boldsymbol{Y}_{Q_2,n-m}))\right] \\
&= \frac{1}{n!}\sum_{\tau \in \mathfrak{S}_n} \mathbb{E}_{\mathbf{Y}_{Q_1,m},\mathbf{Y}_{Q_2,m}}\left[f(\tau \star (\boldsymbol{Y}_{Q_1,m}, \boldsymbol{Y}_{Q_2,n-m}))\right] \\
&= \frac{1}{n!}\sum_{\substack{I \subset [\![1,n]\!] \\ |I|=m}} \sum_{\sigma_1 \in \mathfrak{S}_m} \sum_{\sigma_2 \in \mathfrak{S}_{n-m}} \mathbb{E}_{\mathbf{Y}_{Q_1,m},\mathbf{Y}_{Q_2,m}}\left[f(\tau_I \star (\sigma_1 \star \boldsymbol{Y}_{Q_1,m}, \sigma_2 \star \boldsymbol{Y}_{Q_2,n-m}))\right] \\
&= \frac{1}{\binom{n}{m}}\sum_{\substack{I \subset [\![1,n]\!] \\ |I|=m}} \mathbb{E}_{\sigma_1,\sigma_2,\mathbf{Y}_{Q_1,m},\mathbf{Y}_{Q_2,m}}\left[f(\tau_I \star (\sigma_1 \star \boldsymbol{Y}_{Q_1,m}, \sigma_2 \star \boldsymbol{Y}_{Q_2,n-m}))\right]
\end{aligned}$$

Now note that by induction, $\sigma_1 \star \boldsymbol{Y}_{Q_1,m} \sim X_{\rho,Q_1}^{\otimes m}$ and $\sigma_2 \star \boldsymbol{Y}_{Q_2,n-m} \sim X_{\rho,Q_2}^{\otimes(n-m)}$.

Let $X_1^1, ..., X_1^n \sim X_{\rho,Q_1}$ and $X_1^1, ..., X_1^n \sim X_{\rho,Q_2}$ be $2n$ i.i.d. random variables; the previous statement shows that , $\tau_I \star (\sigma_1 \star \boldsymbol{Y}_{Q_1,m}, \sigma_2 \star \boldsymbol{Y}_{Q_2,n-m}) \sim (X_1^i \mathbf{1}_{i\in I} + (\mathbf{1} - \mathbf{1}_{i\in I})X_2^i)_{1\leqslant i\leqslant n}$ (here, I is fixed). Moreover, note that $P(M=m) = \binom{n}{m} q^m (1-q)^{n-m}$ where $q = I(Q_1)/I(Q)$. Hence

$$\mathbb{E}[f(\sigma \star \boldsymbol{Y}_{Q,n})] = \sum_{I \subset [\![1,n]\!]} q^{|I|}(1-q)^{n-|I|}\mathbb{E}_{X_1^i,X_2^i}(X_1^i \mathbf{1}_{i\in I} + (\mathbf{1} - \mathbf{1}_{i\in I})X_2^i)_{1\leqslant i\leqslant n}$$

Now let $B_1, ..., B_n$ be n i.i.d. Bernoulli variables of parameter q independent of the X_1, X_2. Note that from the previous equation,

$$\mathbb{E}[f(\sigma \star \boldsymbol{Y}_{Q,n})] = \mathbb{E}[f((X_1^i B_i + X_2^i(1-B_i))_{1\leqslant i\leqslant n})]$$

It is easy to see that $(X_1^i B_i + X_2^i(1-B_i))_{1\leqslant i\leqslant n}$ are i.i.d. and distributed as $X_{\rho,Q}$, which concludes the proof.

□

Proof of Theorem 1. Theorem 1 is now a simple consequence of Lemmas 6 to 8.The bound on the number of integral computations can be easily obtained by noting that for any sample, at most $sum_{k=1}^{d} q_k^\rho$ hyper-rectangles are visited (we do not count the first since this computation is done once and for all in any case). Since $q_k^\rho = \lceil \log_2(\delta_k/\rho) \rceil \leqslant \log_2(2\delta_k/\rho)$, this yields a bound of $\log_2(2^d|Q|/\rho^d) = \log_2(|Q|) + d\log_2(2/\rho)$ per sample, hence the result. □

D.3 Evaluating the error of the sampling algorithm : proof of Theorem 2

Theorem 2 is a specific case of the following theorem. For a given function g defined on a hyper-rectangle Q, define its Lipschitz constant with respect to the infinity norm :

$$\forall x \in Q,\ \|x\|_\infty = \sup_{1 \leqslant k \leqslant d} |x_i|, \qquad \mathrm{Lip}_\infty(g) = \sup_{\substack{x,y \in Q \\ x \neq y}} \frac{|g(x) - g(y)|}{\|x - y\|_\infty}. \tag{68}$$

Theorem 7 (Variation bounds)**.** *Let Q be a hyper-rectangle, $\rho > 0$, $p_Q = f/I(Q)$ and $p_{Q,\rho}$ defined in Eq. (7). Recall the definition of $\mathrm{Lip}_\infty(f), \mathrm{Lip}_\infty(\sqrt{f})$ from Eq. (68). The following bounds hold.*

$$d_{TV}(p_Q, p_{Q,\rho}) \leqslant \tfrac{|Q|}{I(Q)}\ \mathrm{Lip}_\infty(f)\ \rho \tag{69}$$

$$H(p_Q, p_{Q,\rho}) \leqslant \sqrt{\tfrac{|Q|}{I(Q)}}\ \mathrm{Lip}_\infty(\sqrt{f})\ \rho \tag{70}$$

$$\mathbb{W}_p(p_Q, p_{Q,\rho}) \leqslant \sqrt{d}\rho, \qquad p \geqslant 1. \tag{71}$$

Proof. Recall that $p_Q = f\mathbf{1}_Q/I(Q)$ and hence

$$\forall x \in Q,\ p_Q(x) = \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} f(x)\mathbf{1}_{Q_\rho}(x)$$

Combining the previous equation with Eq. (7), it holds :

$$\forall x \in Q,\ p_Q(x) - p_{Q,\rho}(x) = \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} (f(x) - \tfrac{I(Q_\rho)}{|Q_\rho|})\mathbf{1}_{Q_\rho}(x) \tag{72}$$

1. Distance between f and its mean on a small cube. Let $Q_\rho \in \mathcal{D}_{Q,\rho}$ and $x \in Q_\rho$, it holds

$$|f(x) - \tfrac{I(Q_\rho)}{|Q_\rho|}| \leqslant \mathrm{Lip}_\infty(f)\ \rho. \tag{73}$$

Indeed, expanding the mean, we get $f(x) - \frac{I(Q_\rho)}{|Q_\rho|} = \frac{1}{|Q_\rho|}\int_{Q_\rho}(f(x) - f(y))\,dy$. Moreover, $|f(x) - f(y)| \leqslant \mathrm{Lip}_\infty(f)\ \|x - y\|_\infty$. Plugging that back in the previous equation and using the fact that $\|x - y\|_\infty \leqslant \rho$ on Q_ρ, we get Eq. (73)

2. Bounds on the total variation and L^2 distances. Using Eqs. (72) and (73), we immediately get

$$\begin{aligned}\int_Q |p_Q(x) - p_{Q,\rho}(x)|dx &= \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \int_{Q_\rho} |f(x) - \tfrac{I(Q_\rho)}{|Q_\rho|}|dx \\ &\leqslant \tfrac{|Q|\,\mathrm{Lip}_\infty(f)\ \rho}{I(Q)}.\end{aligned}$$

3. Bound on the Wasserstein norm $\mathbb{W}_p$. Consider the following density on $Q \times Q$:

$$\gamma(x,y) = \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} f(x)\mathbf{1}_{Q_\rho}(x)\tfrac{1}{|Q_\rho|}\mathbf{1}_{Q_\rho}(y). \tag{74}$$

A simple computation shows that $\gamma \in \Pi(p_Q, p_{Q,\rho})$ (see Santambrogio [17] and Eq. (33)), i.e. that its marginals are p_Q and $p_{Q,\rho}$. Hence, by definition Eq. (33), we have

$$\mathbb{W}_p^p(p_Q, p_{Q,\rho}) \leqslant \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \int_{Q_\rho \times Q_\rho} |x-y|^p \tfrac{f(x)}{|Q_\rho|}\, dxdy.$$

Now using the fact that if $x, y \in Q_\rho$, we have $\|x-y\| \leqslant \sqrt{d}\rho$ as Q_ρ is a hyper-rectangle with all sides of length less than or equal to ρ, we finally get : $\mathbb{W}_p(p_Q, p_{Q,\rho}) \leqslant \sqrt{d}\rho$

4. Hellinger distance bound. Note that we could get a looser bound using Eq. (36) which only relies on the Lipschitz constant of f and not on that of $\sqrt{f}$. Here, we concentrate on that case.

Let $Q_\rho \in \mathcal{D}_{Q,\rho}$. By the intermediate value theorem, there exists $z \in Q_\rho$ such that $f(z) = \frac{I(Q_\rho)}{|Q_\rho|}$ and hence for any $x \in Q_\rho$, it holds

$$\left|\sqrt{f}(x) - \sqrt{\tfrac{I(Q_\rho)}{|Q_\rho|}}\right| = \left|\sqrt{f}(x) - \sqrt{f}(z)\right| \leqslant \mathrm{Lip}_\infty(\sqrt{f})\, \|x-z\|_\infty \leqslant \mathrm{Lip}_\infty(\sqrt{f})\, \rho.$$

Bounding the distance between $p_{Q,\rho}$ and p_Q by decomposing on dyadic hyper-rectangles using the previous expression, it holds

$$\begin{aligned}
H(p_Q, p_{Q,\rho})^2 &= \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \int_{Q_\rho} \left|\sqrt{\tfrac{f(x)}{I(Q)}} - \sqrt{\tfrac{I(Q_\rho)}{|Q_\rho| I(Q)}}\right|^2 dx \\
&= \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \int_{Q_\rho} \left|\sqrt{f(x)} - \sqrt{\tfrac{I(Q_\rho)}{|Q_\rho|}}\right|^2 dx \\
&\leqslant \tfrac{(\mathrm{Lip}_\infty(\sqrt{f})\ \rho)^2}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \int_{Q_\rho} 1\, dx = \left(\sqrt{\tfrac{|Q|}{I(Q)}}\, \mathrm{Lip}_\infty(\sqrt{f})\, \rho\right)^2.
\end{aligned}$$

□

D.4 Time complexity

In the Theorem 1, we measure the cost of the algorithm in terms of evaluation of integrals of the PSD model and in particular in the number of calls to the erf function (or subtractions) in the computation of such integrals. The fact that this is the true bottleneck of the algorithm can be seen in Appendix D.4, as integrals take 95% of the CPU time.

E A general method of approximation and sampling

In this section, we prove proposition 1 and Theorem 3 using mainly results from Rudi and Ciliberto [15]. We introduce those results sequentially, showing the how each one is a building block towards the final result.

Table 1: Main computing times (% of the CPU time)

PART	MAIN OPERATION	TIME
Integration	Eqs. (3) to (5)	
Computing $K_{X,\eta/2}$		71%
Computing $\overline{X}$		6%
Computing $G_{X,2\eta,Q}$	Computing $\mathcal{A}, \mathcal{B}$	8%
	Calls to erf	6%
	Other	8%
Other		1%
Sampling	algorithm 1	
Computing $I(Q)$	Calls to erf	34%
	Computing $\mathcal{A}, \mathcal{B}$	26%
	Mulitplications $\sqrt{\eta}$	11%
	Other	24%
Other		5%

For this section, fix a probability distribution p on the set $\mathcal{X} = (-1,1)^d$ (this is for the sake of simplicity; any hyper-rectangle could do), and assume that Assumption 1 holds for a certain $\beta \in \mathbb{N},\ \beta > 0$, i.e. there exists $J \in \mathbb{N}$ and $q_1, ..., q_J \in W_2^\beta(\mathcal{X}) \cap L^\infty(\mathcal{X})$ such that $p = \sum_j q_j^2$. In this section, this probability distribution p is only known through a function f_p proportional to its density. Denote with $Z_p > 0$ this proportionality constant, i.e. $f_p/Z_p = p$, and with f_j the renormalized $q_j : q_j/\sqrt{Z_p} = f_j$ s.t. $f_p = \sum_j f_j^2$. Our goal is to be able to generate i.i.d. samples from a distribution as close as possible to p.

To do so, we first approximate f_p by a Gaussian PSD model $\widehat{f}_{\tau,m,\lambda} = f(\cdot; \widehat{A}_{\tau,m,\lambda}, \widetilde{X}_m, \eta)$ where $\eta = \tau \mathbf{1}_d$ and $\tau > 0$, $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ is obtained as $(\tilde{x}_1, ..., \tilde{x}_m)^\top$ from m i.i.d. uniform samples from $\mathcal{X}$, and $\widehat{A}_{\tau,m,\lambda}$ is obtained by solving the problem Eq. (15) which we rewrite here for a given $\lambda > 0$:

$$\min_{A\in\mathbb{S}_+(\mathbb{R}^m)} \int_{\mathcal{X}} f(x; A, X, \eta)^2 dx - 2\sum_{i=1}^{n} f_p(x_i) f(x_i; A, X, \eta) + \lambda \|K^{1/2} A K^{1/2}\|_F, \tag{15}$$

where $K = K_{\widetilde{X}_m,\eta}$ and the $(x_i)_{1\leqslant i\leqslant n}$ represented by $X \in \mathbb{R}^{n\times d}$ are n i.i.d. samples from the uniform distribution on $\mathcal{X}$.

The parameters τ, m, n, λ are selected in order to have an ε approximation of the probability p.

Using the fact that we can easily compute integrals of Gaussian PSD models, we can easily have access to $\widehat{p}_{\tau,m,\lambda} = \widehat{f}_{\tau,m,\lambda}/\widehat{Z}_{\tau,m,\lambda}$ where $\widehat{Z}_{\tau,m,\lambda} = \|\widehat{f}_{\tau,m,\lambda}\|_{L^1(\mathcal{X})} = \int_{\mathcal{X}} \widehat{f}_{\tau,m,\lambda}(x)\, dx$.

We then apply algorithm 1 to $\widehat{p}_{\tau,m,\lambda}$, the hyper-rectangle $\mathcal{X}$, the desired number of samples N and a certain ρ controlling the size of the dyadic decomposition of $\mathcal{X}$ in order to sample from a distribution whose total variation distance to p is less than a constant times ε.

Existence of a compressed ε-close Gaussian PSD model. We start by invoking Theorem 6 in order to obtain an ε-approximation of f_p in the form of a general PSD $f_{\tau,\epsilon}$ with associated operator $M_{\tau,\epsilon} \in \mathbb{S}_+(\mathcal{H}_\eta)$. This PSD model can then be compressed using a compression operator as described in Appendix C.2. This is the object of the following proposition.

Proposition 8 (Compression of $M_{\tau,\epsilon}$)**.** *Let $\varepsilon \in (0,1]$, $\tau \geqslant \varepsilon^{-2/\beta}$ and define $\eta = \tau \mathbf{1}_d \in \mathbb{R}^d$. Let $M_{\tau,\epsilon}$ be given by Theorem 6 applied to f_p and satisfying Eq. (63) and $f_{\tau,\epsilon}$ the corresponding PSD model.*

Let $m \in \mathbb{N}$, $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ be a data matrix corresponding to vectors $\tilde{x}_1, ..., \tilde{x}_m$ which are sampled independently and uniformly from $\mathcal{X}$, and $\widetilde{P}_{\eta,m}$ be the associated orthogonal projection in $\mathcal{H}_\eta$. Let $\widetilde{M}_{\tau,m,\epsilon} := \widetilde{P}_{\eta,m} M_{\tau,\epsilon} \widetilde{P}_{\eta,m}$ be the operator associated to the compressed PSD model $\widetilde{f}_{\tau,m,\epsilon}$ of $f_{\tau,\epsilon}$ onto $\widetilde{X}_m$ (see Eq. (49) and Eq. (50) for the definitions).

Let $\delta \in (0,1]$. If one of the two following are true

$$m \geqslant C_1' \tau^{d/2} \left(\log \tfrac{C_2'}{\varepsilon} + \tfrac{d}{2}\log\tau\right)^d \left(\log \tfrac{C_3'}{\delta} + \tfrac{d}{2}\log\tau + \log\log \tfrac{C_2'}{\varepsilon}\right); \tag{75}$$

$$m \geqslant C_1'' \varepsilon^{-d/\beta} \left(\log \tfrac{C_2'}{\varepsilon}\right)^d \left(\log \tfrac{C_2'}{\varepsilon} + \log \tfrac{C_3'}{\delta}\right), \qquad \tau = \varepsilon^{-2/\beta} \tag{76}$$

then with probability at least $1-\delta$, it holds

$$\begin{aligned} \|f_{\tau,\epsilon} - \widetilde{f}_{\tau,m,\epsilon}\|_{L^2(\mathcal{X})} \leqslant 2^d \|f_{\tau,\epsilon} - \widetilde{f}_{\tau,m,\epsilon}\|_{L^\infty(\mathcal{X})} \leqslant 2C \ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\, \varepsilon \\ \operatorname{Tr}\left(\widetilde{M}_{\tau,m,\epsilon}\right) \leqslant \operatorname{Tr}\left(M_{\tau,\epsilon}\right) \leqslant C \ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta} \ \tau^{d/2} \end{aligned} \tag{77}$$

The constants $C, C_1', C_2', C_3', C_1''$ depends only on d, β, and not on $\tau, \varepsilon, m, \delta$.

Proof. Using Eq. (63) in Theorem 6 applied to f_p, we see that if $\varepsilon \leqslant 1$ and $\tau \geqslant \varepsilon^{-2/\beta}$, there exists constants C_4, C_5 depending only on d ,β, and not on τ, ε such that $\|f(\cdot; M_{\tau,\epsilon}, \phi_\eta) - f_p\|_{L^2(\mathcal{X})} \leqslant C_4 \ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ \varepsilon$ and $\operatorname{Tr}(M_{\tau,\epsilon}) \leqslant C_5 \ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ \tau^{d/2}$ (we set $C_5 = C_2(1+e^{C_3})$ where C_2, C_3 are introduced in Theorem 6). Now setting $\rho = \frac{\varepsilon}{2^d\, \tau^{d/2}}$ which is less than 1 since $\varepsilon \leqslant 1$ and $\tau \geqslant \varepsilon^{-2/\beta} \geqslant 1$, we can apply Lemma 3 and hence, with probability at least $1-\delta$, if

$$m \geqslant C_1 \tau^{d/2} (\log \tfrac{C_2\, \tau^{d/2}}{\varepsilon})^d \left(\log \tfrac{C_3}{\delta} + \log\tau + \log\log \tfrac{C_2\, \tau^{d/2}}{\varepsilon}\right), \tag{78}$$

with $C_1 \leftarrow C_1$ from Lemma 3, $C_2 \leftarrow \max(e, C_2 2^d)$ where C_2 is given by Lemma 3 and $C_3 \leftarrow C_3$ from Lemma 3, it holds $\sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| \leqslant \rho$ (hence C_1, C_2, C_3 depend only on d).

1. Let us now show that Eq. (78) is implied by Eq. (75). Let us bound :

$$\begin{aligned} \log\log \tfrac{C_2\, \tau^{d/2}}{\varepsilon} &= \log\left(\log \tfrac{C_2}{\varepsilon}\left(1 + \frac{d/2\log\tau}{\log \frac{C_2}{\varepsilon}}\right)\right) \\ &= \log\log \tfrac{C_2}{\varepsilon} + \log\left(1 + \frac{d/2\log\tau}{\log \frac{C_2}{\varepsilon}}\right) \\ &\leqslant \log\log \tfrac{C_2}{\varepsilon} + \tfrac{d}{2}\log\tau, \end{aligned}$$

where the last inequality is obtained since $\log(1+t) \leqslant t$ and $C_2/\varepsilon \geqslant C_2 \geqslant e$ by definition of C_2 and since $\varepsilon \leqslant 1$. Setting $C_1' = 3C_1$, $C_2' = C_2$ and $C_3' = C_3$, it is therefore clear that Eq. (75) implies Eq. (78).

2. Moreover, Eq. (75) is in turn implied by Eq. (76). Indeed, in the case where $\tau = \varepsilon^{-2/\beta}$, we have the bound

$$\log\log \tfrac{C_2'}{\varepsilon} + \tfrac{d}{2}\log\tau \leqslant \log \tfrac{C_2'}{\varepsilon} + \tfrac{d}{2}\log\tau = \log \tfrac{C_2'}{\varepsilon} + \tfrac{d}{\beta}\log\tfrac{1}{\varepsilon} \leqslant (1 + d/\beta)\log\tfrac{C_2'}{\varepsilon}$$

since $C_2' \geqslant e \geqslant 1$. Thus, taking $C_1'' = C_1'(1+d/\beta)^{d+1}$, Eq. (76) implies Eq. (75).

3. If Eq. (78) holds, then Eq. (77) holds with probability at least $1-\delta$. Indeed, for the first part, since Eq. (78) holds, with probability at least $1-\delta$, $\sup_{x\in\mathcal{X}} \|(I-\widetilde{P}_{\eta,m})\phi_\eta(x)\| \leqslant \rho = \frac{\varepsilon}{2^d \tau^{d/2}}$

Moreover, using Eq. (61) combined with the fact that for any $x \in \mathcal{X}$, $|f(x; M_{\tau,\epsilon}, \phi_\eta)| = |\langle \phi_\eta(x), M_{\tau,\epsilon}\phi_\eta(x)\rangle| \leqslant \|\phi_\eta(x)\|^2_{\mathcal{H}_\eta}\|M_{\tau,\epsilon}\| = \|M_{\tau,\epsilon}\|$ since $\|\phi_\eta(x)\|^2 = k_\eta(x,x) = 1$, it holds

$$\|f(\cdot; M_{\tau,\epsilon}, \phi_\eta) - f(\cdot; \widetilde{M}_{\tau,m,\epsilon}, \phi_\eta)\|_{L^\infty(\mathcal{X})} \leqslant \|M_{\tau,\epsilon}\|(\rho^2+\rho) \leqslant 2\|M_{\tau,\epsilon}\|\rho.$$

We conclude using the fact that for any operator M, and any orthogonal projection P, $\|M\| \leqslant \mathrm{Tr}\,(M)$ and $\mathrm{Tr}\,(PMP) \leqslant \mathrm{Tr}\,(M)$. We then conclude the proof by using the definition of ρ and the fact that $\int_{\mathcal{X}} 1\, dx = 2^d$, and setting $C \leftarrow C_5$.

□

Combining Eq. (63) and Eq. (77), we see that if m is large enough, one can find a Gaussian PSD model of the form $\widetilde{f}_{\tau,m,\epsilon} = f(\,\cdot\,; \widetilde{A}_{\tau,m,\epsilon}, \widetilde{X}_m, \tau \mathbf{1}_d)$ (where $\widetilde{A}_{\tau,m,\epsilon}$ is defined through Eq. (50) from $\widetilde{M}_{\tau,m,\epsilon}$) which is $C\,\|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\,\varepsilon$ close to f_p and whose trace is controlled. It now remains to compare the performance of $\widetilde{f}_{\tau,m,\epsilon}$ with the Gaussian PSD model learned from evaluations of f_p, $\widehat{f}_{\tau,m,\lambda}$, which is the solution of Eq. (15) which we can compute.

Controlling the L^2 distance between $\widehat{f}_{\tau,m,\lambda}$ and f_p. This theorem is a rewriting of Theorem 7 of Rudi and Ciliberto [15], but with the point of view of ε instead of n.

Proposition 9 (Performance of $\widehat{f}_{\tau,m,\lambda}$)**.** *Let $n \in \mathbb{N}$ and let $(x_1, ..., x_n)$ be n i.i.d. samples from p. Let $\delta \in (0,1]$ and $\varepsilon \leqslant \frac{1}{e}$. Assume n satisfies*

$$n \geqslant \varepsilon^{-(d+2\beta)/\beta} \log^d\left(\tfrac{1}{\varepsilon}\right) \log\left(\tfrac{2}{\delta}\right), \tag{79}$$

Let $m \in \mathbb{N}$ and assume m satisfies Eq. (76) and let $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ be a data matrix corresponding to vectors $\tilde{x}_1, ..., \tilde{x}_m$ which are sampled independently and uniformly from $\mathcal{X}$. Let $\lambda = \varepsilon^{2(\beta+d)/\beta}$, $\tau = \varepsilon^{-2/\beta}$ and $\widehat{f}_{\tau,m,\lambda}$ be the Gaussian PSD model associated to the solution $\widehat{A}_{\tau,m,\lambda}$ of Eq. (15) with $\widetilde{X}_m, \lambda, \tau$. With probability at least $1-2\delta$, the following holds

$$\left(\|\widehat{f}_{\tau,m,\lambda} - f_p\|^2_{L^2(\mathcal{X})} + \lambda\|\widehat{M}_{\tau,m,\lambda}\|_F^2\right)^{1/2} \leqslant C\ \|f_p\|_{\mathsf{sos},\mathcal{X},\beta}\ \varepsilon, \tag{80}$$

where C is a constant depending only on d, β, and not on $\varepsilon, \delta, \lambda, m, \tau, f_p$.

Proof. We start by applying the same reasoning as in the proof of Theorem 7 by Rudi and Ciliberto [15].

Note that since $\tau = \varepsilon^{-2/\beta}$ and Eq. (76) is satisfied, with probability at least $1-\delta$, it holds $\|\widetilde{f}_{\tau,m,\epsilon} - \widehat{f}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})} \leqslant 2C_1\ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ \varepsilon$ (where $C_1 \leftarrow C$ from Eq. (77)) and hence $\|f_p - \widehat{f}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})} \leqslant (C_0 + 2C_1)\ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ \varepsilon$, (where $C_0 \leftarrow C_1$ from Theorem 6). C_0, C_1 are both constants depending only on d, β. Moreover, since the Frobenius norm is bounded by the trace norm, by definition of τ, we also have $\|\widehat{M}_{\tau,m,\lambda}\|_F \leqslant \mathrm{Tr}\,(\widehat{M}_{\tau,m,\lambda}) \leqslant C_1\ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ \tau^{d/2} \leqslant C_1\ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ \varepsilon^{-d/\beta}$.

We can modify Theorem E.2 from Rudi and Ciliberto [15] by taking $\widehat{v} = \frac{1}{n}\sum_{i=1}^n f_p(x_i)\psi_\eta(x_i)$ and $v = \int_{\mathcal{X}} f_p(x)\psi_\eta(x)\, dx$; all the formulas then remain true and adapt to our problem Eq. (15). Applying Theorem E.2 from Rudi and Ciliberto [15] to $\widetilde{A}_{\tau,m,\epsilon}$ and using Lemma E.3 of Rudi and Ciliberto [15] to simplify notation, as well as the bound on the term $\|Q_\lambda^{-1/2}(\widehat{v} - v)\|$

combining Lemma E.4 (with $\zeta = Q_\lambda^{-1/2} f_p(x)\psi_\eta(x)$) using $s = d$ and Lemma E.5 (again, for more details, see part 2 of the proof of Theorem 7 by Rudi and Ciliberto [15]) and using the fact that $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b},\ a, b \geqslant 0$, with probability at least $1-\delta$, it holds :

$$\begin{aligned}\left(\|\widehat{f}_{\tau,m,\lambda} - f_p\|^2_{L^2(\mathcal{X})} + \lambda\|\widehat{M}_{\tau,m,\lambda}\|_F^2\right)^{1/2} &\leqslant \|\widetilde{f}_{\tau,m,\epsilon} - f_p\|_{L^2(\mathcal{X})} \\ &+ \sqrt{\lambda}\|\widetilde{M}_{\tau,m,\epsilon}\|_F + C_2\ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ \frac{\log\frac{2}{\delta}}{n\lambda^{1/4}} \\ &+ C_3\ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ \frac{\tau^{d/4}\left(\log\frac{1}{\lambda}\right)^{d/2}\left(\log\frac{2}{\delta}\right)^{1/2}}{n^{1/2}}, \quad (81)\end{aligned}$$

where C_2 and C_3 are constants which depend only on d.

Note that in the proof of Lemma E.4 of Rudi and Ciliberto [15], $\|\zeta\|$ is bounded in essential supremum and standard deviation by $\|f_p\|_{L^\infty(\mathcal{X})}\times$ a quantity independent of f_p which is then bounded, hence the previous concentration bound since $\|f_p\|_{L^\infty(\mathcal{X})} \leqslant \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}$.

Now combining both events in a union bound, and plugging in the fact that $\lambda = \varepsilon^{\frac{2\beta+2d}{\beta}}$ and $\tau = \varepsilon^{-2/\beta}$, we see that with probability at least $1-2\delta$, the left hand term is bounded by the following quantity:

$$\varepsilon\ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}\ (C_0 + 3C_1 + T), \quad (82)$$

$$T = C_2\frac{\varepsilon^{-\frac{3\beta+d}{2\beta}}\log\frac{2}{\delta}}{n} + C_3\frac{\varepsilon^{-(d+2\beta)/2\beta}\left(\frac{2\beta+2d}{\beta}\log\frac{1}{\varepsilon}\right)^{d/2}\left(\log\frac{2}{\delta}\right)^{1/2}}{n^{1/2}}.$$

Now the goal is to bound the term T. Note that as soon as $\varepsilon \leqslant e^{-1}$ and $\delta \leqslant 2$, if $Y = \frac{\varepsilon^{-(d+2\beta)/\beta}\log^d\left(\frac{1}{\varepsilon}\right)\log\left(\frac{2}{\delta}\right)}{n}$, then it holds $T \leqslant \frac{C_2}{\log^2 2}Y + C_3\sqrt{Y}$. Now note that $Y \leqslant 1$ iif $n \geqslant \varepsilon^{-(d+2\beta)/\beta}\log^d\left(\frac{1}{\varepsilon}\right)\log\left(\frac{2}{\delta}\right)$. The theorem therefore holds with $C \leftarrow 1 + 3C_1 + C_2/\log^2 2 + C_3$.

Finally, the fact that all bounds involving $\|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta}$ can be replaced, up to constants depending only on β, d, by the norm $\|f_p\|_{\mathsf{sos},\mathcal{X},\beta}$, is simply a consequence of proposition 4. □

We now come to the final part of our section detailing the proof of proposition 10 and Theorem 8, which consists in approximately sampling from the learnt model $\widehat{f}_{\tau,m,\lambda}$ using algorithm 1 with well chosen parameters.

Performance of the re-normalized probability measure $\widehat{p}_{\tau,m,\lambda}$. We start off with a technical lemma.

Lemma 9 (Technical lemma). *Let $\|\cdot\|$ be a norm on a vector space E, and let $x, y \in E \setminus \{0\}$. Then it holds:*

$$\left\|\frac{a}{\|a\|} - \frac{b}{\|b\|}\right\| \leqslant \frac{2\|a-b\|}{\|a\|}. \quad (83)$$

Moreover, if $\|a-b\| \leqslant \|a\|/2$, it holds

$$\frac{\|a\|}{\|b\|} \leqslant 2. \quad (84)$$

Proof. Introduce the quantity $\frac{b}{\|a\|}$ in order to get

$$\left\|\frac{a}{\|a\|} - \frac{b}{\|b\|}\right\| \leqslant \left\|\frac{a}{\|a\|} - \frac{b}{\|a\|}\right\| + \left\|\frac{b}{\|a\|} - \frac{b}{\|b\|}\right\| = \frac{\|a-b\|}{\|a\|} + \|b\|\left|\frac{1}{\|a\|} - \frac{1}{\|b\|}\right|.$$

One concludes by writing

$$\left|\frac{1}{\|a\|}-\frac{1}{\|b\|}\right| = \frac{|\|b\|-\|a\||}{\|a\|\,\|b\|} \leqslant \frac{\|b-a\|}{\|a\|\,\|b\|},$$

where the last inequality is simply the triangle inequality. This concludes the proof of Eq. (83). The proof of Eq. (84) is simply the result of applying the bound $\frac{1}{\|b\|} \leqslant \frac{1}{\|a\|-\|b-a\|} \leqslant \frac{2}{\|a\|}$. □

Proposition 10 (Performance of $\widehat{p}_{\tau,m,\lambda}$)**.** *Let p be a probability density w.r.t. the Lebesgue measure on $\mathcal{X}=(-1,1)^d$ satisfying Assumption 1 for a certain β. There exists $\varepsilon_0>0$ depending only on d,β, and $\|p\|_{\mathsf{sos},\mathcal{X},\beta}$ and C_1,C_2,C_1',C_2',C_3' depending only on d,β such that the following holds.*

Let $n\in\mathbb{N}$ and let $(x_1,...,x_n)$ be n i.i.d. samples selected uniformly at random from $\mathcal{X}$. Let $\delta\in(0,1]$ and $\varepsilon\leqslant\varepsilon_0$, $\lambda=\varepsilon^{2(\beta+d)/\beta}$ and $\tau=\varepsilon^{-2/\beta}$. Assume n satisfies Eq. (79), i.e.

$$n \geqslant \varepsilon^{-(d+2\beta)/\beta}\log^d\left(\tfrac{1}{\varepsilon}\right)\log\left(\tfrac{2}{\delta}\right). \tag{79}$$

Let $m\in\mathbb{N}$ and assume m satisfies Eq. (76), i.e.

$$m \geqslant C_1'\varepsilon^{-d/\beta}\left(\log\tfrac{C_2'}{\varepsilon}\right)^d\left(\log\tfrac{C_2'}{\varepsilon}+\log\tfrac{C_3'}{\delta}\right), \tag{76}$$

and let $\widetilde{X}_m\in\mathbb{R}^{m\times d}$ be a data matrix corresponding to vectors $\tilde{x}_1,...,\tilde{x}_m$ which are sampled independently and uniformly from $\mathcal{X}$.

Let $\widehat{f}_{\tau,m,\lambda}$ be the Gaussian PSD model associated to the solution $\widehat{A}_{\tau,m,\lambda}$ of Eq. (15) with $\widetilde{X}_m,\lambda,\tau$ and let $\widehat{p}_{\tau,m,\lambda}$ be the associated probability density on $\mathcal{X}$ (i.e. the re-normalization of $\widehat{f}_{\tau,m,\lambda}$). Let $\widehat{R}_{\tau,m,\lambda}$ be PSD operator on $\mathcal{H}_\eta$ associated to $\widehat{p}_{\tau,m,\lambda}$. With probability at least $1-2\delta$, it holds

$$d_{TV}(\widehat{p}_{\tau,m,\lambda},p) \leqslant C_1\ \|p\|_{\mathsf{sos},\mathcal{X},\beta}\ \varepsilon, \qquad \|\widehat{R}_{\tau,m,\lambda}\|_F \leqslant C_2\ \|p\|_{\mathsf{sos},\mathcal{X},\beta}\ \varepsilon^{-d/\beta}. \tag{85}$$

Proof. Since the assumptions of proposition 9 are satisfied, we have by Eq. (80) the existence of a constant C depending only on d, β, and not on $\varepsilon,\delta,\lambda,m,\tau,f_p$, such that

$$\|\widehat{f}_{\tau,m,\lambda}-f_p\|_{L^2(\mathcal{X})} \leqslant C\ \|f_p\|_{\mathsf{sos},\mathcal{X},\beta}\ \varepsilon, \qquad \|\widehat{M}_{\tau,m,\lambda}\|_F \leqslant C\ \|f_p\|_{\mathsf{sos},\mathcal{X},\beta}\ \varepsilon^{-d/\beta}, \tag{86}$$

where we have used the fact that $\lambda=\varepsilon^{2+2d/\beta}$.

Now using the fact that $\|\bullet\|_{L^1(\mathcal{X})}\leqslant 2^{d/2}\|\bullet\|_{L^2(\mathcal{X})}$ (by Cauchy-Schwarz inequality), Eq. (86) shows in particular that $\|\widehat{f}_{\tau,m,\lambda}-f_p\|_{L^1(\mathcal{X})}\leqslant 2^{d/2}C\ \|f_p\|_{\mathsf{sos},\mathcal{X},\beta}\ \varepsilon$. Now applying Eq. (83) of Lemma 9, using the fact that $\widehat{p}_{\tau,m,\lambda}=\widehat{f}_{\tau,m,\lambda}/\|\widehat{f}_{\tau,m,\lambda}\|_{L^1(\mathcal{X})}$ and $p=f_p/\|f_p\|_{L^1(\mathcal{X})}$, it holds

$$\begin{aligned} d_{TV}(\widehat{p}_{\tau,m,\lambda},p) = \|\widehat{p}_{\tau,m,\lambda}-p\|_{L^1(\mathcal{X})} &\leqslant 2\|\widehat{f}_{\tau,m,\lambda}-f_p\|_{L^1(\mathcal{X})}/\|f_p\|_{L^1(\mathcal{X})} \\ &\leqslant 2^{d/2+1}C\ \|f_p\|_{\mathsf{sos},\mathcal{X},\beta}/\|f_p\|_{L^1(\mathcal{X})}\ \varepsilon. \end{aligned} \tag{87}$$

Since $p=f_p/\|f_p\|_{L^1(\mathcal{X})}$, we have $\|f_p\|_{\mathsf{sos},\mathcal{X},\beta}/\|f_p\|_{L^1(\mathcal{X})}=\|p\|_{\mathsf{sos},\mathcal{X},\beta}$. This shows

$$d_{TV}(\widehat{p}_{\tau,m,\lambda},p) \leqslant 2^{d/2+1}C\ \|p\|_{\mathsf{sos},\mathcal{X},\beta}\,\varepsilon.$$

Now set $\varepsilon_0=\min(e^{-1},2^{-d/2-1}C^{-1}\|p\|^{-1}_{\mathsf{sos},\mathcal{X},\beta})$. If $\varepsilon\leqslant\varepsilon_0$, we have $2^{d/2}C\,\|f_p\|_{\mathsf{sos},\mathcal{X},\beta}\,\varepsilon\leqslant\|f_p\|_{L^1(\mathcal{X})}/2$ and hence $\|\widehat{f}_{\tau,m,\lambda}-f_p\|_{L^1(\mathcal{X})}\leqslant\|f_p\|_{L^1(\mathcal{X})}/2$. By Eq. (84) of Lemma 9, we therefore have $\|f_p\|_{L^1(\mathcal{X})}/\|\widehat{f}_{\tau,m,\lambda}\|_{L^1(\mathcal{X})}=Z_p/\widehat{Z}_{\tau,m,\lambda}\leqslant 2$. Now since $\widehat{R}_{\tau,m,\lambda}=\widehat{M}_{\tau,m,\lambda}/\widehat{Z}_{\tau,m,\lambda}$, using Eq. (86), it holds $\|\widehat{R}_{\tau,m,\lambda}\|\leqslant C_2\ \|p\|_{\mathsf{sos},\mathcal{X},\beta}\ \varepsilon^{-d/\beta}$ where $C_2=2C$, which depends only on β,d.

□

Theorem 8 (Performance of $p_{\texttt{sample}}$)**.** *Under the assumptions and notations of the previous theorem (proposition 1), there exists a constant C_3 depending only on d, β, such that the following holds.*

Let $\widehat{p}_{\tau,m,\lambda}$ be given by the previous proposition. Let $p_{\texttt{sample}}$ be the dyadic approximation of $\widehat{p}_{\tau,m,\lambda}$ on $Q = \mathcal{X} = (-1,1)^d$ and of width ρ (see Eq. (7)). Recall from Theorem 1 that algorithm 1 applied to $Q = (-1,1)^d, N, \rho$ returns N i.i.d. samples from $p_{\texttt{sample}}$.

If on the one hand ρ is set to $\varepsilon^{1+(d+1)/\beta}$, then with probability at least $1-2\delta$,

$$d_{TV}(\widehat{p}_{\tau,m,\lambda}, p_{\texttt{sample}}) \leqslant C_3 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon, \qquad d_{TV}(p, p_{\texttt{sample}}) \leqslant (C_1 + C_3) \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon. \tag{88}$$

If on the other ρ is set adaptively to guarantee $d_{TV}(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leqslant \varepsilon$ as in Remark 1 then with probability at least $1-2\delta$, $\rho \geqslant \varepsilon^{1+(d+1)/\beta}/(C_3 \ \|p\|_{\texttt{sos},\mathcal{X},\beta})$, and hence

$$d_{TV}(\widehat{p}_{\tau,m,\lambda}, p_{\texttt{sample}}) \leqslant \varepsilon, \qquad d_{TV}(p, p_{\texttt{sample}}) \leqslant C_1 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon + \varepsilon. \tag{89}$$

In any case, this guarantees that the complexity in terms of erf *computations is bounded by*

$$O(Nm^2 \log \tfrac{1}{\rho}) = O\left(N \ \varepsilon^{-2d/\beta} \log^{2d+1}\left(\tfrac{1}{\varepsilon}\right)\left(\log\left(\tfrac{1}{\varepsilon}\right) + \log\left(\tfrac{1}{\delta}\right)\right)\right), \tag{90}$$

where the O notations is taken with constants depending on $d, \beta, \|p\|_{\texttt{sos},\mathcal{X},\beta}$.

Proof. Let us bound $\mathrm{Lip}_\infty(\widehat{p}_{\tau,m,\lambda})$. Note that

$$\mathrm{Lip}_\infty(\widehat{p}_{\tau,m,\lambda}) \leqslant \sup_{x\in\mathcal{X}} \sum_{k=1}^{d} \partial_k \widehat{p}_{\tau,m,\lambda}(x).$$

Using Lemma 5, we get $\mathrm{Lip}_\infty(\widehat{p}_{\tau,m,\lambda}) \leqslant d 2^{3/2}\sqrt{\tau}\|\widehat{R}_{\tau,m,\lambda}\|$. Using the fact that $\tau = \varepsilon^{-2/\beta}$ and that by Eq. (85), $\|\widehat{R}_{\tau,m,\lambda}\| \leqslant \|\widehat{R}_{\tau,m,\lambda}\|_F \leqslant C_2 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon^{-d/\beta}$, we therefore have $\mathrm{Lip}_\infty(\widehat{p}_{\tau,m,\lambda}) \leqslant 2^{3/2} d C_2 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon^{-(d+1)/\beta}$. Hence, applying Theorem 7 to $\widehat{p}_{\tau,m,\lambda}$, we get

$$d_{TV}(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leqslant 2^{3/2} \ 2^d d C_2 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon^{-(d+1)/\beta} \ \rho. \tag{91}$$

On the one hand, if we use algorithm 1 with $\rho = \varepsilon^{1+\frac{(d+1)}{\beta}}$, by the previous equation, we get $d_{TV}(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leqslant 2^{3/2} d \ 2^d C_2 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon$.

If on the other hand we find ρ adaptively by computing a bound

$$\widetilde{\mathrm{Lip}}(A) = 2^{3/2}\tau^{1/2} d \|K^{1/2} A K^{1/2}\| = 2^{3/2}\tau^{1/2} d \|\widehat{R}_{\tau,m,\lambda}\|_F$$

from $\widehat{p}_{\tau,m,\lambda}$ as in Remark 1, and finding ρ such that $2^d \widetilde{\mathrm{Lip}}(A) \ \rho = \frac{|Q|}{I(Q)} \widetilde{\mathrm{Lip}}(A) \ \rho = \varepsilon$, since the adaptive bound will have computed

$$\widetilde{\mathrm{Lip}}(A) \ \leqslant 2^{3/2} d C_2 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon^{-(d+1)/\beta},$$

we will get $\rho \geqslant \frac{\varepsilon^{1+(d+1)/\beta}}{2^{d+3/2} \ d \ C_2 \ \|p\|_{\texttt{sos},\mathcal{X},\beta}}$ and hence $d_{TV}(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leqslant \varepsilon$. The last point is just a consequence of Theorem 1 and the bound on m in Eq. (76). □

F Approximation and sampling using a rank one PSD model

In this section, we prove the results in Sec. 4.2, i.e. proposition 2 and Theorem 4.

For this section, fix a probability which has density p with respect to the Lebesgue measure dx on $\mathcal{X} = (-1, 1)^d$, (this is for the sake of simplicity; any hyper-rectangle could do), and assume that Assumption 2 holds for a certain $\beta \in \mathbb{N},\ \beta > 0$, i.e. there exists $q \in W_2^\beta(\mathcal{X}) \cap L^\infty(\mathcal{X})$ such that $p = q^2$. This is the case, for instance, when $p \propto e^{-V(x)}$ where V is β times differentiable.

One of the main advantages of our method will be to deal with probability measures which are known up to a constant; therefore, in this section, we take f_p such that $p = f_p/Z(f_p)$ where $Z(f_p) = \int_{\mathcal{X}} f_p(x)dx$. Assuming Assumption 2 holds, we take $g_p \in W_2^\beta(\mathcal{X}) \cap L^\infty(\mathcal{X})$ such that $g_p^2 = f_p$ as and assume that p is only known through function evaluations of g_p, i.e. we can evaluate the function $g_p(x)$ for any $x \in \mathcal{X}$.

Once again, our goal is to be able to generate N i.i.d. samples from a distribution which is ε-close to p, in a sense which we will define. To do so, we first approximate g_p by a Gaussian linear model $\widehat{g}_{\tau,m,\lambda} = g(\bullet;\ \widehat{a}_{\tau,m,\lambda}, \widetilde{X}_m, \eta)$ (see Eq. (2) for a definition) where $\eta = \tau \mathbf{1}_d$ for some $\tau > 0$, $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ is obtained as $(\tilde{x}_1, ..., \tilde{x}_m)^\top$ from m i.i.d. uniform samples from $\mathcal{X}$, and $\widehat{a}_{\tau,m,\lambda}$ is obtained by solving the problem Eq. (20) which we rewrite here for a given $\lambda > 0$ and for n i.i.d. samples $(x_1, .., x_n)$ sampled uniformly from $\mathcal{X}$:

$$\widehat{a}_{\tau,m,\lambda} = \operatorname*{argmin}_{a\in\mathbb{R}^m} \frac{1}{n}\sum_{i=1}^{n} \left(g(x_i;\ a, \tilde{x}_m, \tau\mathbf{1}_d) - g_p(x_i)\right)^2 + \lambda a^\top K_{\widetilde{X}_m,\eta} a. \tag{20}$$

This yields a Gaussian linear model $\widehat{g}_{\tau,m,\lambda} \in \mathcal{H}_\eta$ of g_p. Since $\widehat{g}^2_{\tau,m,\lambda} = \widehat{f}_{\tau,m,\lambda}$ is a PSD model (indeed $\widehat{f}_{\tau,m,\lambda} = f(\bullet;\ \widehat{A}_{\tau,m,\lambda}, \widetilde{X}_m, \tau\mathbf{1}_d)$ with $\widehat{A}_{\tau,m,\lambda} = \widehat{a}_{\tau,m,\lambda}\widehat{a}^\top_{\tau,m,\lambda}$), we can see $\widehat{f}_{\tau,m,\lambda}$ as a Gaussian PSD model of f_p, and hence its renormalized version $\widehat{p}_{\tau,m,\lambda}$ as a PSD model of p.

The parameters τ, m, λ, n are selected in order to have an ε approximation of the probability p.

Furthermore, note that the first term in the optimized quantity in Eq. (20) is an empirical version of the quantity

$$\frac{1}{|\mathcal{X}|}\int_{\mathcal{X}} \left|\sqrt{\widehat{f}_{\tau,m,\lambda}(x)} - \sqrt{f_p}(x)\right|^2 dx \leqslant \frac{1}{|\mathcal{X}|}\int_{\mathcal{X}} |\widehat{g}_{\tau,m,\lambda}(x) - g_p(x)|^2\ dx.$$

This quantity is related to Hellinger distance $H(p, \widehat{p}_{\tau,m,\lambda})$ defined in Eq. (32).

This will therefore be the natural measure in which to express the quality of the approximation $\widehat{p}_{\tau,m,\lambda}$ of p in this section.

The bound obtained on the performance of $\widehat{p}_{\tau,m,\lambda}$ can be decomposed into two steps.

- We start by bounding the distance between any $g \in \mathcal{H}_\eta$ and $\widehat{g}_{\tau,m,\lambda}$ in Theorem 9.
- We then select a $g_{\tau,\varepsilon}$ which is ε-close to g_p, and use it as a reference point in order to bound the distance between g_p and $\widehat{g}_{\tau,m,\lambda}$. To do so, we need to apply different concentration inequalities to obtain a final bound in terms of performance for both $\widehat{f}_{\tau,m,\lambda}$ with respect to f_p and $\widehat{p}_{\tau,m,\lambda}$ with respect to p in Hellinger distance in proposition 2.

Bound on the performance of $\widehat{g}_{\tau,m,\lambda}$ compared to an arbitrary function g. Here, we adapt Theorem 2. from Rudi et al. [13].

Theorem 9 (Bounding the error [13]). *Let $\eta \in \mathbb{R}^d_{++}$ and $g \in \mathcal{H}_\eta$.*

$$\|C_{\eta,\lambda}^{1/2}(g - \widehat{g}_{\tau,m,\lambda})\| \leqslant \theta_1^2\theta_2 \, \|\widehat{g}_p - \widehat{S}_\eta g\|_{\mathbb{R}^n} + \|g\|_{\mathcal{H}_\eta}(1 + \theta_1\theta_2 + \theta_1^2) \left(\sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| + \lambda^{1/2}\right), \tag{92}$$

where $\theta_1 = \|\widehat{C}_{\eta,\lambda}^{-1/2} C_{\eta,\lambda}^{1/2}\|$, $\theta_2 = \|\widehat{C}_{\eta,\lambda}^{1/2} C_{\eta,\lambda}^{-1/2}\|$ and $\widehat{g}_p = (g_p(x_i)/\sqrt{n})_{1\leqslant i\leqslant n} \in \mathbb{R}^n$.

Proof. Let $g \in \mathcal{H}_\eta$. We can apply a modification of Theorem 2 by Rudi et al. [13]. Indeed, consider in the notations of Rudi et al. [13] the loss $\mathcal{E}(f) = \|C_\eta^{1/2}(f - g)\|_{\mathcal{H}_\eta}$, and note that the assumptions are satisfied with $\nu = 0$ and $R = \|g\|_{\mathcal{H}_\eta}$, since g minimizes $\mathcal{E}$ and $\|C^{-0}g\|_{\mathcal{H}_\eta} = \|g\|_{\mathcal{H}_\eta}$. Moreover, note that in the proof of that theorem, one can replace C_η by $C_{\eta,\lambda}$ without changing the result (indeed, in the proof, one always bounds $\|C_\eta^{1/2} \star\| \leqslant \|C_\eta^{1/2} C_{\eta,\lambda}^{-1/2}\| \, \|C_{\eta,\lambda}^{1/2} \star\| \leqslant \|C_{\eta,\lambda}^{1/2} \star\|$). Thus, in that setting, without combining the "constant" terms in the bounds and looking into the proof of Theorem 2 of Rudi et al. [13], it holds

$$\|C_{\eta,\lambda}^{1/2}(\widehat{g}_{\tau,m,\lambda} - g)\| \leqslant \theta_1^2 \, \|C_{\eta,\lambda}^{-1/2} \widehat{S}_\eta^*(\widehat{g}_p - \widehat{S}_\eta g)\| + R(1 + \theta_1\theta_2)\|(I - \widetilde{P}_{\eta,m}) C_{\eta,\lambda}^{1/2}\| + R\theta_1^2\lambda^{1/2}, \tag{93}$$

where $\theta_1 = \|\widehat{C}_{\eta,\lambda}^{-1/2} C_{\eta,\lambda}^{1/2}\|$ and $\theta_2 = \|\widehat{C}_{\eta,\lambda}^{1/2} C_{\eta,\lambda}^{-1/2}\|$.

Note that $\|C_{\eta,\lambda}^{-1/2} \widehat{S}_\eta^*(\widehat{g}_p - \widehat{S}_\eta g)\| \leqslant \|C_{\eta,\lambda}^{-1/2} \widehat{S}_\eta^*\| \, \|\widehat{g}_p - \widehat{S}_\eta g\|_{\mathbb{R}^n} \leqslant \theta_2 \, \|\widehat{g}_p - \widehat{S}_\eta g\|_{\mathbb{R}^n}$ since $\|C_{\eta,\lambda}^{-1/2} \widehat{S}_\eta^*\|^2 = \|C_{\eta,\lambda}^{-1/2} \widehat{C}_\eta C_{\eta,\lambda}^{-1/2}\| \leqslant \|C_{\eta,\lambda}^{-1/2} \widehat{C}_{\eta,\lambda} C_{\eta,\lambda}^{-1/2}\| = \theta_2^2$.

Moreover, using the definition of C_η, it holds

$$\begin{aligned}
\|(I - \widetilde{P}_{\eta,m}) C_{\eta,\lambda}^{1/2}\|^2 &= \|(I - \widetilde{P}_{\eta,m}) C_\eta (I - \widetilde{P}_{\eta,m}) + \lambda(I - \widetilde{P}_{\eta,m})\| \\
&\leqslant \tfrac{1}{|X|} \left\| \int_{\mathcal{X}} (I - \widetilde{P}_{\eta,m})\phi_\eta(x) \otimes \phi_\eta(x)(I - \widetilde{P}_{\eta,m}) \, dx \right\| + \lambda \|(I - \widetilde{P}_{\eta,m})\| \\
&\leqslant \sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\|^2 + \lambda.
\end{aligned}$$

Combining these results and using the fact that $\sqrt{a + b} \leqslant \sqrt{a} + \sqrt{b}$ for any $a, b \geqslant 0$, we get the bound. □

Performance of $\widehat{p}_{\tau,m,\lambda}$. We can now state the main results of this section, i.e. the bound on the performance of $\widehat{p}_{\tau,m,\lambda}$.

Proposition 11 (Performance of $\widehat{p}_{\tau,m,\lambda}$). *Let p be a probability density on $\mathcal{X} = (-1, 1)^d$, and assume $p = q^2$ and $q \in L^\infty(\mathcal{X}) \cap W_2^\beta(\mathcal{X})$ for some $\beta \geqslant 0$. Let $\widetilde{\nu} > \min(1, d/(2\beta))$. There exists a constant ε_0 depending only on $\|q\|_{L^\infty(\mathcal{X})}, \|q\|_{W_2^\beta(\mathcal{X})}, \beta, d$, constants C_1, C_2, C_3, C_4, C_5 depending only on β, d and a constant C_1' depending only on $\beta, d, \widetilde{\nu}$ such that the following holds.*

Let $\delta \in (0, 1]$ and $\varepsilon \leqslant \varepsilon_0$, and assume $(x_1, ..., x_n)$ and $(\tilde{x}_1, ..., \tilde{x}_m)$ are respectively n and m uniform i.i.d. samples on $\mathcal{X}$, satisfying

$$m \geqslant C_1 \varepsilon^{-d/\beta} \log^d \tfrac{C_2}{\varepsilon} \log \tfrac{C_3}{\delta\varepsilon} \tag{94}$$

$$n \geqslant C_1' \varepsilon^{-2\widetilde{\nu}} \log \tfrac{8}{\delta} \tag{95}$$

Let $\tau = \varepsilon^{-2/\beta}$, $\eta = \tau \mathbf{1}_d$ and $\lambda = \varepsilon^{2+d/\beta}$. Let $\widehat{a}_{\tau,m,\lambda} \in \mathbb{R}^n$ be the vector obtained by solving Eq. (20) and $\widehat{g}_{\tau,m,\lambda} \in \mathcal{H}_\eta$ the associated Gaussian linear model (see Eq. (2)). Let $\widehat{f}_{\tau,m,\lambda} = \widehat{g}^2_{\tau,m,\lambda}$ be the associated Gaussian PSD model, $\widehat{Z}_{\tau,m,\lambda} = \int_{\mathcal{X}} \widehat{f}_{\tau,m,\lambda}(x)\, dx$ be the normalizing constant, and $\widehat{p}_{\tau,m,\lambda} = \widehat{f}_{\tau,m,\lambda}/\widehat{Z}_{\tau,m,\lambda}$ be the renormalized PSD model, which is a probability density. Let $\widehat{R}_{\tau,m,\lambda}$ be PSD operator in $\mathbb{S}_+(\mathcal{H}_\eta)$ associated to $\widehat{p}_{\tau,m,\lambda}$.

With probability at least $1 - 3\delta$, it holds

$$
\begin{aligned}
H(\widehat{p}_{\tau,m,\lambda}, p) &\leqslant C_4 \|q\|_{L^\infty(\mathcal{X}) \cap W_2^\beta(\mathcal{X})} \varepsilon \\
\operatorname{Tr}\left(\widehat{R}_{\tau,m,\lambda}\right) &= \left\| \frac{\widehat{g}_{\tau,m,\lambda}}{\sqrt{\widehat{Z}_{\tau,m,\lambda}}} \right\|^2_{\mathcal{H}_\eta} \leqslant C_5 \|q\|^2_{L^\infty(\mathcal{X}) \cap W_2^\beta(\mathcal{X})} \varepsilon^{-d/\beta},
\end{aligned} \tag{96}
$$

where $\| \bullet \|_{L^\infty(\mathcal{X}) \cap W_2^\beta(\mathcal{X})} = \max(\| \bullet \|_{W_2^\beta(\mathcal{X})}, \| \bullet \|_{L^\infty(\mathcal{X})})$.

Proof. Let $\tau > 0$, and define $\eta = \tau \mathbf{1}_d$. By proposition 4, we can extend g_p to the whole of $\mathbb{R}^d$ and there exists an constant C such that $\|g_p\|_{W_2^\beta(\mathbb{R}^d)} \leqslant \|g_p\|_{W_2^\beta(\mathcal{X})}$ and $\|g_p\|_{L^\infty(\mathbb{R}^d)} \leqslant C \|g_p\|_{L^\infty(\mathcal{X})}$. We still denote with g_p such an extension. Let $g_{\tau,\varepsilon}$ be given by proposition 7 when approximating g_p.

Setting $\tau = \varepsilon^{-2/\beta}$ and $\lambda = \varepsilon^{\frac{2\beta+d}{\beta}}$, since we assume $\varepsilon \leqslant 1$, Eq. (51) gives us two constants C_1, C_2 depending only on β, d such that

$$
\begin{cases} \|g_{\tau,\varepsilon} - g_p\|_{L^2(\mathbb{R}^d)} \leqslant \varepsilon \|g_p\|_{W_2^\beta(\mathbb{R}^d)} \\ \|g_{\tau,\varepsilon} - g_p\|_{L^\infty(\mathbb{R}^d)} \leqslant C_1\, \varepsilon^{1-\nu}\, \|g_p\|_\bullet \end{cases} \|g_{\tau,\varepsilon}\|_{\mathcal{H}_\eta} \leqslant C_2\, \|g_p\|_{W_2^\beta(\mathbb{R}^d)}\, \tau^{d/4} = C_2\, \|g_p\|_{W_2^\beta(\mathbb{R}^d)} \varepsilon^{-\frac{d}{2\beta}}.
$$

1. Bounding $\|\widehat{g}_p - \widehat{S}_\eta g_{\tau,\varepsilon}\|_{\mathbb{R}^n}$ Apply Theorem 3 of Boucheron et al. [3], reformulated in Proposition 10 from Rudi et al. [13]. Consider the random variable $\zeta = (g_{\tau,\varepsilon} - g_p)(X)^2 - \frac{1}{|\mathcal{X}|}\|g_{\tau,\varepsilon} - g_p\|^2_{L^2(\mathcal{X})}$ where X follows the uniform law on $\mathcal{X}$. Then $|\zeta| \leqslant \|g_{\tau,\varepsilon} - g_p\|^2_{L^\infty(\mathcal{X})}$ almost surely, and $\mathbb{E}[\zeta^2] \leqslant \|g_{\tau,\varepsilon} - g_p\|^2_{L^\infty(\mathcal{X})} \frac{1}{|\mathcal{X}|}\|g_{\tau,\varepsilon} - g_p\|^2_{L^2(\mathcal{X})}$. Applying the concentration bound yields that with probability at least $1 - \delta$, it holds

$$
\begin{aligned}
\|\widehat{g}_p - \widehat{S}_\eta g_{\tau,\varepsilon}\|^2_{\mathbb{R}^n} - \tfrac{1}{|\mathcal{X}|}\|g_{\tau,\varepsilon} - g_p\|^2_{L^2(\mathcal{X})} \leqslant{}& \frac{2\|g_{\tau,\varepsilon} - g_p\|^2_{L^\infty(\mathcal{X})} \log\frac{1}{\delta}}{3n} \\
&+ \sqrt{\frac{2\|g_{\tau,\varepsilon} - g_p\|^2_{L^\infty(\mathcal{X})} \frac{1}{|\mathcal{X}|}\|g_{\tau,\varepsilon} - g_p\|^2_{L^2(\mathcal{X})} \log\frac{1}{\delta}}{n}},
\end{aligned}
$$

and thus

$$
\|\widehat{g}_p - \widehat{S}_\eta g_{\tau,\varepsilon}\|^2_{\mathbb{R}^n} \leqslant \left(\frac{1}{\sqrt{|\mathcal{X}|}} \|g_{\tau,\varepsilon} - g_p\|_{L^2(\mathcal{X})} + \|g_{\tau,\varepsilon} - g_p\|_{L^\infty(\mathcal{X})} \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \right)^2.
$$

Hence, by Eq. (51), and because $|\mathcal{X}| = 2^d$, there exists two constants C_3 and C_4 depending only on d and β such that with probability at least $1 - \delta$, it holds

$$
\|\widehat{g}_p - \widehat{S}_\eta g_{\tau,\varepsilon}\|_{\mathbb{R}^n} \leqslant C_3 \varepsilon\, \|g_p\|_{W_2^\beta(\mathbb{R}^d)} + C_4 \varepsilon \frac{\|g_p\|_\bullet \log\frac{1}{\delta}}{\varepsilon^\nu \sqrt{n}}. \tag{97}
$$

2. Guaranteeing $\sup_{x\in\mathcal{X}} \|(I-\widetilde{P}_{\eta,m})\phi_\eta(x)\| \leqslant \lambda^{1/2} = \varepsilon^{1+d/(2\beta)}$ Using Lemma 3 and proceeding in the same way as in point 2 of the proof of proposition 8, we see that there exists constants C_5, C_6, C_7 depending only on d and β such that as soon as

$$m \geqslant C_5 \varepsilon^{-d/\beta} \left(\log \tfrac{C_6}{\varepsilon}\right)^d \log \tfrac{C_7}{\delta\varepsilon}, \tag{98}$$

it holds $\sup_{x\in\mathcal{X}} \|(I-\widetilde{P}_{\eta,m})\phi_\eta(x)\| \leqslant \lambda^{1/2}$ with probability at least $1-\delta$.

3. Finding a lower bound for $\|C_\eta\|$ This will be necessary in the next bound. Let $v(z) = k_\eta(0,z) = e^{-\tau\|z\|^2}$. Then $\|v\|_{\mathcal{H}_\eta} = 1$ and

$$\begin{aligned}
\|C_\eta^{1/2} v\|_{\mathcal{H}}^2 &= \tfrac{1}{|\mathcal{X}|}\int_{\mathcal{X}} |v(x)|^2\, dx \\
&= \tfrac{1}{|\mathcal{X}|}\left(\int_{-1}^{1} e^{-2\tau t^2}\, dt\right)^d \\
&\geqslant \tfrac{1}{2^d}\left(\int_{-1}^{1} e^{-2t^2}\, dt\right)^d \tau^{-d/2} = C_8\, \tau^{-d/2},
\end{aligned}$$

where the last inequality comes from the fact that $\tau \geqslant 1$ since $\varepsilon \leqslant 1$. Hence, $\|C_\eta\| \geqslant C_8 \tau^{-d/2}$ where C_8 is a constant depending only on d. Hence, as soon as $\lambda \leqslant C_8\tau^{-d/2}$ which rewrites $\varepsilon \leqslant \sqrt{C_8}$, it holds $\lambda \leqslant \|C_\eta\|$.

4. Bounding θ_1, θ_2**.** Using the same reasoning as that of Proposition 2. of Rudi et al. [13], if $b = \|C_{\eta,\lambda}^{-1/2}(\widehat{C}_\eta - C_\eta)C_{\eta,\lambda}^{-1/2}\|$, then $\theta_1 \leqslant 1/(1-b)$ and $\theta_2^2 \leqslant 1+b$. Bounding b can be done using Proposition 8 of Rudi et al. [13]: if $\lambda \leqslant \|C_\eta\|$, and $\delta \in (0,1]$ it holds, with probability at least $1-\delta$:

$$\|C_{\eta,\lambda}^{-1/2}(\widehat{C}_\eta - C_\eta)C_{\eta,\lambda}^{-1/2}\| \leqslant \frac{2(1+\mathcal{N}_\infty(\lambda))\log\frac{8}{\lambda\delta}}{3n} + \sqrt{\frac{2\mathcal{N}_\infty(\lambda)\log\frac{8}{\lambda\delta}}{n}}, \tag{99}$$

where we have used the fact that $\mathrm{Tr}\,(C_\eta) \leqslant 1$.

Note that $\mathcal{N}_\infty(\lambda) = \sup_{x\in\mathcal{X}} \|C_{\eta,\lambda}^{-1/2}\phi_\eta(x)\|^2 \leqslant C_9 \tau^{(s-d)d/(2s)}\lambda^{-d/(2s)}$ for any $s > d/2$ where C_9 depends only on s, d by a proof completely analogous as that of Step 2 of Lemma E.4 by Rudi and Ciliberto [15]. Replacing the values of τ, λ yields : $\mathcal{N}_\infty(\lambda) \leqslant C_9 \varepsilon^{-\frac{2d(\beta+s)-d^2}{2s\beta}}$.

Note that the function $\gamma : s \in]d/2, +\infty[\mapsto \frac{2d\beta+2ds-d^2}{2s\beta}$ is a homography and therefore reaches all the values $\widetilde{\nu}$ strictly between 2 and d/β.

Therefore, for any $\widetilde{\nu} > \nu$, there exists a constant C_{10} depending only on d and $\widetilde{\nu}$ such that $(1+\mathcal{N}_\infty(\lambda))\log\frac{1}{\lambda} \leqslant C_{10}\varepsilon^{-2\widetilde{\nu}}$.

Hence, there exists a constant depending only on $d, \beta, \widetilde{\nu}$ such that if $n \geqslant C_{11}\varepsilon^{-2\widetilde{\nu}}\log\frac{8}{\delta}$, and if $\varepsilon \leqslant \min(1/2, \sqrt{C_8})$ then $b \leqslant 1/3$ (here we have bounded $\log\frac{8}{\delta\lambda}$ by a constant times $\log\frac{1}{\lambda}\log\frac{8}{\delta}$ provided $\varepsilon \leqslant 1/2$ and hence $\lambda \leqslant 1/4$. Moreover, note that C_{11} can be taken large enough, by Eq. (97), to guarantee the following, also with probability $1-\delta$:

$$\|\widehat{g}_p - \widehat{S}_\eta g_{\tau,\varepsilon}\|_{\mathbb{R}^n} \leqslant C_3\varepsilon\, \|g_p\|_{W_2^\beta(\mathbb{R}^d)} + C_4\varepsilon\|g_p\|_\bullet. \tag{100}$$

5. Applying Theorem 9 to $g_{\tau,\varepsilon}$. Combining all the previous equations, we get that if $n \geqslant C_{11}\varepsilon^{-2\widetilde{\nu}} \log \frac{8}{\delta}$, $\varepsilon \leqslant \min(1/2, \sqrt{C_8})$ and $m \geqslant C_5 \varepsilon^{-d/\beta} \left(\log \frac{C_6}{\varepsilon}\right)^d \log \frac{C_7}{\delta\varepsilon}$, it holds Eq. (100) and $b \leqslant 1/3$ as well as $\sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| \leqslant \lambda^{1/2}$ and hence, using the bound on $g_{\tau,\varepsilon}$, there exists a constant C_{12} depending only on d, β such that

$$\|C_{\eta,\lambda}^{1/2}(g_{\tau,\varepsilon} - \widehat{g}_{\tau,m,\lambda})\| \leqslant C_{12} \max(\|g_p\|_{W_2^\beta(\mathbb{R}^d)}, \|g_p\|_\bullet)\varepsilon.$$

Thus, using the bound on $\|g_{\tau,\varepsilon} - g_p\|_{L^2(\mathbb{R}^d)}$, and the fact that $gC_\eta g = \frac{1}{|\mathcal{X}|}\|g\|^2_{L^2(\mathcal{X})}$ we get

$$\begin{aligned} \|g_p - \widehat{g}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})} &\leqslant C_{13} \max(\|g_p\|_{W_2^\beta(\mathbb{R}^d)}, \|g_p\|_\bullet)\varepsilon, \\ \|\widehat{g}_{\tau,m,\lambda}\|_{\mathcal{H}_\eta} &\leqslant C_{14} \max(\|g_p\|_{W_2^\beta(\mathbb{R}^d)}, \|g_p\|_\bullet)\, \varepsilon^{-d/2\beta}. \end{aligned} \tag{101}$$

6. Bounding the performance of $\widehat{p}_{\tau,m,\lambda}$. Note that $q = \frac{g_p}{\|g_p\|_{L^2(\mathcal{X})}}$ and $\sqrt{\widehat{p}_{\tau,m,\lambda}} = \frac{|\widehat{g}_{\tau,m,\lambda}|}{\|\widehat{g}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})}}$. Thus, using Eq. (83), it holds

$$\begin{aligned} H(\widehat{p}_{\tau,m,\lambda}, p) &= \left\| \frac{g_p}{\|g_p\|_{L^2(\mathcal{X})}} - \frac{|\widehat{g}_{\tau,m,\lambda}|}{\|\widehat{g}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})}} \right\|_{L^2(\mathcal{X})} \\ &\leqslant 2\frac{\|\widehat{g}_{\tau,m,\lambda} - g_p\|_{L^2(\mathcal{X})}}{\|g_p\|_{L^2(\mathcal{X})}}. \end{aligned}$$

Hence, since $q = g_p/\|g_p\|_{L^2(\mathcal{X})}$, we have by Eq. (101) :

$$H(p, \widehat{p}_{\tau,m,\lambda}) \leqslant 2C_{13} \max(\|q\|_{W_2^\beta(\mathbb{R}^d)}, \|q\|_\bullet)\varepsilon.$$

Moreover, by Eq. (84), if $2C_{13} \max(\|q\|_{W_2^\beta(\mathbb{R}^d)}, \|q\|_\bullet)\varepsilon \leqslant 1$, then $\frac{\|g_{\tau,\varepsilon}\|_{L^2(\mathcal{X})}}{\|\widehat{g}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})}} \leqslant 2$ and hence again by Eq. (101), $\|\widehat{p}_{\tau,m,\lambda}\|_{\mathcal{H}_\eta} \leqslant 2C_{14} \max(\|q\|_{W_2^\beta(\mathbb{R}^d)}, \|q\|_\bullet)\, \varepsilon^{-d/2\beta}$. Setting $\varepsilon_0 = \min(1/2, \sqrt{C_8}, (2C_{13} \max(\|q\|_{W_2^\beta(\mathbb{R}^d)}, \|q\|_\bullet))^{-1})$, we therefore have all the desired properties.

7. Replacing norms on $\mathbb{R}^d$ with norm on $\mathcal{X}$. To do so, we just use proposition 4, which does not change anything up to multiplicative constants depending only on d, β.

□

Theorem 10 (Performance of $p_{\texttt{sample}}$)**.** *Under the assumptions and notations of the previous theorem (proposition 11), there exists a constant C_6 depending only on d, β, such that the following holds. Let $\widehat{p}_{\tau,m,\lambda}$ be given by the previous proposition. Let $p_{\texttt{sample}}$ be the dyadic approximation of $\widehat{p}_{\tau,m,\lambda}$ on $Q = \mathcal{X} = (-1,1)^d$ and of width ρ (see Eq. (7)). Recall from Theorem 1 that algorithm 1 applied to $Q = (-1,1)^d, N, \rho$ returns N i.i.d. samples from $p_{\texttt{sample}}$.*

If on the one hand ρ is set to $\varepsilon^{1+(d+2)/(2\beta)}$, then with probability at least $1 - 3\delta$,

$$\begin{aligned} H(\widehat{p}_{\tau,m,\lambda}, p_{\texttt{sample}}) &\leqslant C_6 \|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}\, \varepsilon, \\ H(p, p_{\texttt{sample}}) &\leqslant (C_4 + C_6)\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}\, \varepsilon. \end{aligned} \tag{102}$$

If on the other ρ is set adaptively to guarantee $H(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leqslant \varepsilon$ as in Remark 1, then with probability at least $1-3\delta$,

$$\rho \geqslant \varepsilon^{1+(d+2)/\beta}/(C_6 \, \|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}),$$
$$H(\widehat{p}_{\tau,m,\lambda}, p_{\texttt{sample}}) \leqslant \varepsilon, H(p, p_{\texttt{sample}}) \leqslant (C_1+1)\varepsilon. \tag{103}$$

In any case, this guarantees that the complexity in terms of erf *computations is bounded by*

$$O(Nm^2 \log \tfrac{1}{\rho}) = O\left(N \, \varepsilon^{-2d/\beta} \log^{2d+1}\left(\tfrac{1}{\varepsilon}\right)\left(\log\left(\tfrac{1}{\varepsilon}\right) + \log\left(\tfrac{1}{\delta}\right)\right)\right), \tag{104}$$

where the O notations is taken with constants depending on d, β, $\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}$.

Proof. Let us bound $\mathrm{Lip}_\infty(\sqrt{\widehat{p}_{\tau,m,\lambda}})$. Note that since for any $x, y \in \mathcal{X}$, it holds

$$\left|\sqrt{\widehat{p}_{\tau,m,\lambda}}(x) - \sqrt{\widehat{p}_{\tau,m,\lambda}}(y)\right| = \left||\widehat{g}_{\tau,m,\lambda}(x)| - |\widehat{g}_{\tau,m,\lambda}(y)|\right|/\sqrt{\widehat{Z}_{\tau,m,\lambda}}$$
$$\leqslant |\widehat{g}_{\tau,m,\lambda}(x) - \widehat{g}_{\tau,m,\lambda}(y)|/\sqrt{\widehat{Z}_{\tau,m,\lambda}},$$

we have $\mathrm{Lip}_\infty(\sqrt{\widehat{p}_{\tau,m,\lambda}}) \leqslant \mathrm{Lip}_\infty(\widehat{g}_{\tau,m,\lambda})/\widehat{Z}_{\tau,m,\lambda}$. Now

$$\mathrm{Lip}_\infty(\widehat{g}_{\tau,m,\lambda}) \leqslant \sup_{x\in\mathcal{X}} \sum_{k=1}^d \partial_k \widehat{g}_{\tau,m,\lambda}(x).$$

Using Lemma 2, we get $\mathrm{Lip}_\infty(\widehat{g}_{\tau,m,\lambda}) \leqslant d\sqrt{2\tau}\|\widehat{g}_{\tau,m,\lambda}\|_{\mathcal{H}_\eta}$. Using the fact that $\tau = \varepsilon^{-2/\beta}$ and that by Eq. (96), $\|\widehat{g}_{\tau,m,\lambda}\|_{\mathcal{H}_\eta}/\sqrt{\widehat{Z}_{\tau,m,\lambda}} \leqslant \sqrt{C_5}\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}\varepsilon^{-d/(2\beta)}$, we therefore have $\mathrm{Lip}_\infty(\sqrt{\widehat{p}_{\tau,m,\lambda}}) \leqslant d\sqrt{2C_5}\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}\varepsilon^{-(d+2)/(2\beta)}$. Hence, applying Theorem 7 to $\widehat{p}_{\tau,m,\lambda}$, we get

$$H(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leqslant 2^{d/2} d\sqrt{2C_5}\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}\varepsilon^{-(d+2)/(2\beta)} \, \rho. \tag{105}$$

On the one hand, if we use algorithm 1 with $\rho = \varepsilon^{1+\frac{(d+2)}{2\beta}}$, by the previous equation, we get $H(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leqslant 2^{d/2} d\sqrt{2C_5}\,\varepsilon$.

If on the other hand we find ρ adaptively by computing an upper bound $\widetilde{\mathrm{Lip}}(a)$ defined in s.t. $\widetilde{\mathrm{Lip}}(a) = \sqrt{2\tau}d\|K^{1/2}a\| = \sqrt{2\tau}d\|\widehat{g}_{\tau,m,\lambda}\|/\sqrt{\widehat{Z}_{\tau,m,\lambda}} \geqslant \mathrm{Lip}_\infty(\sqrt{\widehat{p}_{\tau,m,\lambda}})$ from $\widehat{p}_{\tau,m,\lambda}$ and finding ρ such that $2^{d/2}\widetilde{\mathrm{Lip}}(a)\ \rho = \varepsilon$, we will get $\rho \geqslant \frac{\varepsilon^{1+(d+2)/(2\beta)}}{2^{d/2}\ d\ \sqrt{2C_5}\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}}$ and hence $H(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leqslant \varepsilon$. The last point is just a consequence of Theorem 1 and the bound on m in Eq. (94).

□

G Additional experimental details

As mentioned in Sec. 5, we report in Fig. 4 an experiment in which we learn the density of the indicator function of $[-1,1]$ using algorithm 4.

Note that this is out of the setting of Theorem 4, as these bounds rely on the regularity of the target density which is not at all the case here.

However, in order to sample approximately from p as a rough approximation, algorithm 4 could be relevant : it shows that we must develop tools which analyse these algorithms beyond notions of regularity, with rougher objectives.

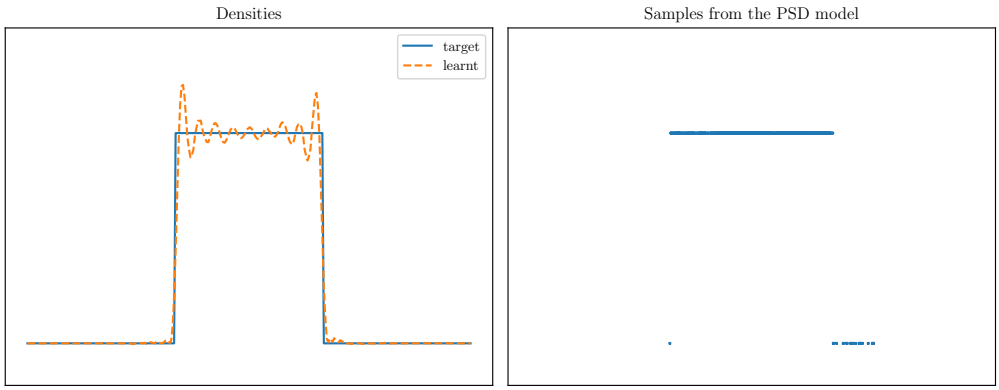

Figure 4: Trying to learn a non-continuous function using a rank one PSD model. *(left)* Plot of the target and learnt distributions using algorithm 4. *(right)* 1000 samples generated from the learnt distribution $p_{\texttt{sample}}$.