Local Optimisation of Nyström Samples Through Stochastic Gradient Descent

Matthew Hutchings(✉) and Bertrand Gauthier

School of Mathematics, Cardiff University Abacws,
Senghennydd Road, Cardiff CF24 4AG, UK
{hutchingsm1,gauthierb}@cardiff.ac.uk

Abstract. We study a relaxed version of the column-sampling problem for the Nyström approximation of kernel matrices, where approximations are defined from multisets of landmark points in the ambient space; such multisets are referred to as Nyström samples. We consider an unweighted variation of the radial squared-kernel discrepancy (SKD) criterion as a surrogate for the classical criteria used to assess the Nyström approximation accuracy; in this setting, we discuss how Nyström samples can be efficiently optimised through stochastic gradient descent. We perform numerical experiments which demonstrate that the local minimisation of the radial SKD yields Nyström samples with improved Nyström approximation accuracy in terms of trace, Frobenius and spectral norms.

Keywords: Low-rank matrix approximation · Nyström method · Reproducing kernel Hilbert spaces · Stochastic gradient descent

1 Introduction

In Data Science, the Nyström method refers to a specific technique for the low-rank approximation of symmetric positive-semidefinite (SPSD) matrices; see e.g. [4,5,10,11,18]. Given an $N \times N$ SPSD matrix $\mathbf{K}$, with $N \in \mathbb{N}$, the Nyström method consists of selecting a sample of $n \in \mathbb{N}$ columns of $\mathbf{K}$, generally with $n \ll N$, and next defining a low-rank approximation $\hat{\mathbf{K}}$ of $\mathbf{K}$ based on this sample of columns. More precisely, let $\boldsymbol{c}_1, \cdots, \boldsymbol{c}_N \in \mathbb{R}^N$ be the columns of $\mathbf{K}$, so that $\mathbf{K} = (\boldsymbol{c}_1 | \cdots | \boldsymbol{c}_N)$, and let $I = \{i_1, \cdots, i_n\} \subseteq \{1, \cdots, N\}$ denote the indices of a sample of n columns of $\mathbf{K}$ (note that I is a multiset, i.e. the indices of some columns might potentially be repeated). Let $\mathbf{C} = (\boldsymbol{c}_{i_1} | \cdots | \boldsymbol{c}_{i_n})$ be the $N \times n$ matrix defined from the considered sample of columns of $\mathbf{K}$, and let $\mathbf{W}$ be the $n \times n$ principal submatrix of $\mathbf{K}$ defined by the indices in I, i.e. the k,l entry of $\mathbf{W}$ is $[\mathbf{K}]_{i_k,i_l}$, the i_k, i_l entry of $\mathbf{K}$. The Nyström approximation of $\mathbf{K}$ defined from the sample of columns indexed by I is given by

$$\hat{\mathbf{K}} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T, \tag{1}$$

with $\mathbf{W}^{\dagger}$ the Moore-Penrose pseudoinverse of $\mathbf{W}$. The column-sampling problem for Nyström approximation consists of designing samples of columns such that the induced approximations are as accurate as possible (see Sect. 1.2 for more details).

G. Nicosia et al. (Eds.): LOD 2022, LNCS 13810, pp. 123–140, 2023.
https://doi.org/10.1007/978-3-031-25599-1_10

1.1 Kernel Matrix Approximation

If the initial SPSD matrix $\mathbf{K}$ is a kernel matrix, defined from a SPSD kernel K and a set or multiset of points $\mathcal{D} = \{x_1, \cdots, x_N\} \subseteq \mathscr{X}$ (and with $\mathscr{X}$ a general ambient space), i.e. the i, j entry of $\mathbf{K}$ is $K(x_i, x_j)$, then a sample of columns of $\mathbf{K}$ is naturally associated with a subset of $\mathcal{D}$; more precisely, a sample of columns $\{\boldsymbol{c}_{i_1}, \cdots, \boldsymbol{c}_{i_n}\}$, indexed by I, naturally defines a multiset $\{x_{i_1}, \cdots, x_{i_n}\} \subseteq \mathcal{D}$, so that the induced Nyström approximation can in this case be regarded as an approximation induced by a subset of points in $\mathcal{D}$. Consequently, in the kernel-matrix framework, instead of relying only on subsets of columns, we may more generally consider Nyström approximations defined from a multiset $\mathcal{S} \subseteq \mathscr{X}$. Using matrix notation, the Nyström approximation of $\mathbf{K}$ defined by a subset $\mathcal{S} = \{s_1, \cdots, s_n\}$ is the $N \times N$ SPSD matrix $\hat{\mathbf{K}}(\mathcal{S})$, with i, j entry

$$\left[\hat{\mathbf{K}}(\mathcal{S})\right]_{i,j} = \mathbf{k}_{\mathcal{S}}^T(x_i)\mathbf{K}_{\mathcal{S}}^{\dagger}\mathbf{k}_{\mathcal{S}}(x_j), \tag{2}$$

where $\mathbf{K}_{\mathcal{S}}$ is the $n \times n$ kernel matrix defined by the kernel K and the subset $\mathcal{S}$, and where

$$\mathbf{k}_{\mathcal{S}}(x) = \big(K(x, s_1), \cdots, K(x, s_n)\big)^T \in \mathbb{R}^n.$$

We refer to such a set or multiset $\mathcal{S}$ as a *Nyström sample*, and to the elements of $\mathcal{S}$ as *landmark points* (the terminology *inducing points* can also be found in the literature); the notation $\hat{\mathbf{K}}(\mathcal{S})$ emphasises that the considered Nyström approximation of $\mathbf{K}$ is induced by $\mathcal{S}$. As in the column-sampling case, the landmark-point-based framework naturally raises questions related to the characterisation and the design of efficient Nyström samples (i.e. samples leading to accurate approximations of $\mathbf{K}$). In this work, for a fixed $n \in \mathbb{N}$, we interpret Nyström samples of size n as elements of $\mathscr{X}^n$, and we investigate the possibility of directly optimising Nyström samples over $\mathscr{X}^n$. We consider the case $\mathscr{X} = \mathbb{R}^d$, with $d \in \mathbb{N}$, but $\mathscr{X}$ may more generally be a differentiable manifold.

Remark 1. Denoting by $\mathcal{H}$ the reproducing kernel Hilbert space (RKHS, see e.g. [1,14]) of real-valued functions on $\mathscr{X}$ associated with K, we may note that the matrix $\hat{\mathbf{K}}(\mathcal{S})$ is the kernel matrix defined by K_S and $\mathcal{D}$, with K_S the reproducing kernel of the closed linear subspace

$$\mathcal{H}_S = \operatorname{span}\{k_{s_1}, \cdots, k_{s_n}\} \subseteq \mathcal{H},$$

where, for $t \in \mathscr{X}$, the function $k_t \in \mathcal{H}$ is defined as $k_t(x) = K(x, t)$, for all $x \in \mathscr{X}$. ◁

1.2 Assessing the Accuracy of Nyström Approximations

In the classical literature on the Nyström approximation of SPSD matrices, the accuracy of the approximation induced by a Nyström sample $\mathcal{S}$ is often assessed through the following criteria:

(C.1) $C_{\mathrm{tr}}(\mathcal{S}) = \left\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\right\|_*$, with $\|.\|_*$ the trace norm;

(C.2) $C_{\mathrm{F}}(\mathcal{S}) = \left\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\right\|_{\mathrm{F}}$, with $\|.\|_{\mathrm{F}}$ the Frobenius norm;
(C.3) $C_{\mathrm{sp}}(\mathcal{S}) = \left\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\right\|_2$, with $\|.\|_2$ the spectral norm.

Although defining relevant and easily interpretable measures of the approximation error, these criteria are relatively costly to evaluate. Indeed, each of them involves the inversion or pseudoinversion of the kernel matrix $\mathbf{K}_{\mathcal{S}}$, with complexity $\mathcal{O}(n^3)$. The evaluation of the criterion (C.1) also involves the computation of the N diagonal entries of $\hat{\mathbf{K}}(\mathcal{S})$, leading to an overall complexity of $\mathcal{O}(n^3 + Nn^2)$. The evaluation of (C.2) involves the full construction of the matrix $\hat{\mathbf{K}}(\mathcal{S})$, with an overall complexity of $\mathcal{O}(n^3 + n^2N^2)$, and the evaluation of (C.3) in addition requires the computation of the largest eigenvalue of an $N \times N$ SPSD matrix, leading to an overall complexity of $\mathcal{O}(n^3 + n^2N^2 + N^3)$. For $\mathscr{X} = \mathbb{R}^d$, the evaluation of the partial derivatives of these criteria (regarded as maps from $\mathscr{X}^n$ to $\mathbb{R}$) with respect to a single coordinate of a landmark point has a complexity similar to the complexity of evaluating the criteria themselves (and there are in this case nd such partial derivatives). Consequently, a direct optimisation of these criteria over $\mathscr{X}^n$ is intractable in most practical applications.

1.3 Radial Squared-Kernel Discrepancy

As a surrogate for the criteria (C.1)–(C.3), and following the connections between the Nyström approximation of SPSD matrices, the approximation of integral operators with SPSD kernels and the kernel embedding of measures, we consider the following *radial squared-kernel discrepancy* criterion (radial SKD, see [7,9]), denoted by R and given by, for $\mathcal{S} = \{s_1, \cdots, s_n\}$,

$$R(\mathcal{S}) = \begin{cases} \|\mathbf{K}\|_{\mathrm{F}}^2 - \frac{1}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2} \left(\sum_{i=1}^{N} \sum_{j=1}^{n} K^2(x_i, s_j) \right)^2, & \text{if } \|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}} > 0, \\ \|\mathbf{K}\|_{\mathrm{F}}^2, & \text{otherwise,} \end{cases} \tag{3}$$

where $K^2(x_i, s_j)$ stands for $\big(K(x_i, s_j)\big)^2$. We may note that $0 \leqslant R(\mathcal{S}) \leqslant \|\mathbf{K}\|_{\mathrm{F}}^2$. In (3), the evaluation of the term $\|\mathbf{K}\|_{\mathrm{F}}^2$ has complexity $\mathcal{O}(N^2)$; nevertheless, this term does not depend on the Nyström sample $\mathcal{S}$, and may thus be regarded as a constant. The complexity of the evaluation of the term $R(\mathcal{S}) - \|\mathbf{K}\|_{\mathrm{F}}^2$, i.e. of the radial SKD up to the constant $\|\mathbf{K}\|_{\mathrm{F}}^2$, is $\mathcal{O}(n^2 + nN)$; for $\mathscr{X} = \mathbb{R}^d$, the same holds for the complexity of the evaluation of the partial derivative of $R(\mathcal{S})$ with respect to a coordinate of a landmark point, see Eq. (5) below. Importantly, and in contrast to the criteria discussed in Sect. 1.2, the evaluation of the radial SKD criterion or of its partial derivatives does not involve the inversion or pseudoinversion of the $n \times n$ matrix $\mathbf{K}_{\mathcal{S}}$.

Remark 2. From a theoretical standpoint, the radial SKD criterion measures the distance, in the Hilbert space of all Hilbert-Schmidt operators on $\mathcal{H}$, between the integral operator corresponding to the initial matrix $\mathbf{K}$ (i.e. the integral operator defined from the kernel K and a uniform measure on $\mathcal{D}$), and the projection of this operator onto the subspace spanned by an integral operator defined from

the kernel K and a uniform measure on $\mathcal{S}$. The radial SKD may also be defined for non-uniform measures, and the criterion in this case depends not only on $\mathcal{S}$, but also on a set of relative weights associated with each landmark point in $\mathcal{S}$; in this work, we only focus on the uniform-weight case. See [7,9] for more details. ◁

The following inequalities hold:

$$\left\|\mathbf{K}-\hat{\mathbf{K}}(\mathcal{S})\right\|_2^2 \leqslant \left\|\mathbf{K}-\hat{\mathbf{K}}(\mathcal{S})\right\|_{\mathrm{F}}^2 \leqslant R(\mathcal{S}) \leqslant \|\mathbf{K}\|_{\mathrm{F}}^2, \text{ and } \frac{1}{N}\left\|\mathbf{K}-\hat{\mathbf{K}}(\mathcal{S})\right\|_*^2 \leqslant \left\|\mathbf{K}-\hat{\mathbf{K}}(\mathcal{S})\right\|_{\mathrm{F}}^2,$$

which, in complement to the theoretical properties enjoyed by the radial SKD, further support the use of the radial SKD as a numerically-affordable surrogate for (C.1)–(C.3) (see also the numerical experiments in Sect. 4).

From now on, we assume that $\mathscr{X} = \mathbb{R}^d$. Let $[s]_l$, with $l \in \{1, \cdots, d\}$, be the l-th coordinate of s in the canonical basis of $\mathbb{R}^d$. For $x \in \mathscr{X}$, we denote by (assuming they exist)

$$\partial^{[\mathrm{l}]}_{[s]_l} K^2(s,x) \quad \text{and} \quad \partial^{[\mathrm{d}]}_{[s]_l} K^2(s,s) \tag{4}$$

the partial derivatives of the maps $s \mapsto K^2(s,x)$ and $s \mapsto K^2(s,s)$ at s and with respect to the l-th coordinate of s, respectively; the notation $\partial^{[\mathrm{l}]}$ indicates that the left entry of the kernel is considered, while $\partial^{[\mathrm{d}]}$ refers to the diagonal of the kernel; we use similar notations for any kernel function on $\mathscr{X} \times \mathscr{X}$.

For a fixed number of landmark points $n \in \mathbb{N}$, the radial SKD criterion can be regarded as a function from $\mathscr{X}^n$ to $\mathbb{R}$. For a Nyström sample $\mathcal{S} = \{s_1, \cdots, s_n\} \in \mathscr{X}^n$, and for $k \in \{1, \cdots, n\}$ and $l \in \{1, \cdots, d\}$, we denote by $\partial_{[s_k]_l} R(\mathcal{S})$ the partial derivative of the map $R : \mathscr{X}^n \to \mathbb{R}$ at $\mathcal{S}$ with respect to the l-th coordinate of the k-th landmark point $s_k \in \mathscr{X}$. We have

$$\begin{aligned}\partial_{[s_k]_l} R(\mathcal{S}) = {} & \frac{1}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^4}\left(\sum_{i=1}^{N}\sum_{j=1}^{n} K^2(s_j,x_i)\right)^2\left(\partial^{[\mathrm{d}]}_{[s_k]_l} K^2(s_k,s_k) + 2\sum_{\substack{j=1,\\ j\neq k}}^{n} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k,s_j)\right) \\ & - \frac{2}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2}\left(\sum_{i=1}^{N}\sum_{j=1}^{n} K^2(s_j,x_i)\right)\left(\sum_{i=1}^{N} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k,x_i)\right).\end{aligned} \tag{5}$$

By mutualising the evaluation of the terms in (5) that do not depend on k and l, the evaluation of the nd partial derivatives of R at $\mathcal{S}$ has a complexity of $\mathcal{O}\big((d+1)(n^2+nN)\big)$; by contrast (and although the pseudoinversion of $\mathbf{K}_{\mathcal{S}}$ can be mutualised), evaluating the nd partial derivatives of the trace criterion has a complexity of $\mathcal{O}\big(d(n^4+n^3N)\big)$.

In this work, we investigate the possibility to use the partial derivatives (5), or stochastic approximations of these derivatives, to directly optimise the radial SKD criterion R over $\mathscr{X}^n$ via gradient or stochastic gradient descent; the stochastic approximation schemes we consider aim at reducing the burden of the

numerical cost induced by the evaluation of the partial derivatives of R when N is large.

The document is organised as follows. In Sect. 2, we discuss the convergence of a gradient descent with fixed step size for the minimisation of R over $\mathscr{X}^n$. The stochastic approximation of the gradient of the radial SKD criterion (3) is discussed in Sect. 3, and some numerical experiments are carried out in Sect. 4. Section 5 consists of a concluding discussion, and the Appendix contains a proof of Theorem 1.

2 A Convergence Result

We use the same notation as in Sect. 1.3 (in particular, we still assume that $\mathscr{X} = \mathbb{R}^d$), and by analogy with (4), for s and $x \in \mathscr{X}$, and for $l \in \{1, \cdots, d\}$, we denote by $\partial^{[\mathrm{r}]}_{[s]_l} K^2(x, s)$ the partial derivative of the map $s \mapsto K^2(x, s)$ with respect to the l-th coordinate of s. Also, for a fixed $n \in \mathbb{N}$, we denote by $\nabla R(\mathcal{S}) \in \mathscr{X}^n = \mathbb{R}^{nd}$ the gradient of $R : \mathscr{X}^n \to \mathbb{R}$ at $\mathcal{S}$; in matrix notation, we have

$$\nabla R(\mathcal{S}) = \Big(\big(\nabla_{s_1} R(\mathcal{S})\big)^T, \cdots, \big(\nabla_{s_1} R(\mathcal{S})\big)^T \Big)^T,$$

with $\nabla_{s_k} R(\mathcal{S}) = \big(\partial_{[s_k]_1} R(\mathcal{S}), \cdots, \partial_{[s_k]_d} R(\mathcal{S})\big)^T \in \mathbb{R}^d$ for $k \in \{1, \cdots, n\}$.

Theorem 1. *We make the following assumptions on the squared-kernel K^2, which we assume hold for all x and $y \in \mathscr{X} = \mathbb{R}^d$, and all l and $l' \in \{1, \cdots, d\}$, uniformly:*

(A.1) there exists $\alpha > 0$ such that $K^2(x, x) \geqslant \alpha$;
(A.2) there exists $M_1 > 0$ such that $\big|\partial^{[\mathrm{d}]}_{[x]_l} K^2(x, x)\big| \leqslant M_1$ and $\big|\partial^{[\mathrm{l}]}_{[x]_l} K^2(x, y)\big| \leqslant M_1$;

(A.3) there exists $M_2 > 0$ such that $\big|\partial^{[\mathrm{d}]}_{[x]_l} \partial^{[\mathrm{d}]}_{[x]_{l'}} K^2(x, x)\big| \leqslant M_2$, $\big|\partial^{[\mathrm{l}]}_{[x]_l} \partial^{[\mathrm{l}]}_{[x]_{l'}} K^2(x, y)\big| \leqslant M_2$ and $\big|\partial^{[\mathrm{l}]}_{[x]_l} \partial^{[\mathrm{r}]}_{[y]_{l'}} K^2(x, y)\big| \leqslant M_2$.

Let $\mathcal{S}$ and $\mathcal{S}' \in \mathbb{R}^{nd}$ be two Nyström samples; under the above assumptions, there exists $L > 0$ such that

$$\big\|\nabla R(\mathcal{S}) - \nabla R(\mathcal{S}')\big\| \leqslant L \big\|\mathcal{S} - \mathcal{S}'\big\|$$

with $\|.\|$ the Euclidean norm of $\mathbb{R}^{nd}$; in other words, the gradient of $R : \mathbb{R}^{nd} \to \mathbb{R}$ is Lipschitz-continuous with Lipschitz constant L.

Since R is bounded from below, for $0 < \gamma \leqslant 1/L$ and independently of the considered initial Nyström sample $\mathcal{S}^{(0)}$, Theorem 1 entails that a gradient descent from $\mathcal{S}^{(0)}$, with fixed stepsize γ for the minimisation of R over $\mathscr{X}^n$, produces a sequence of iterates that converges to a critical point of R. Barring some specific and largely pathological cases, the resulting critical point is likely to be a local minimum of R, see for instance [12]. See the Appendix for a proof of Theorem 1.

The conditions considered in Theorem 1 ensure the existence of a general Lipschitz constant L for the gradient of R; they, for instance, hold for all sufficiently regular Matérn kernels (thus including the Gaussian, or squared-exponential, kernel). These conditions are only sufficient conditions for the convergence of a gradient descent for the minimisation of R. By introducing additional problem-dependent conditions, some convergence results might be obtained for more general squared kernels K^2 and adequate initial Nyström samples $\mathcal{S}^{(0)}$. For instance, the condition (A.1) simply aims at ensuring that $\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2 \geqslant n\alpha > 0$ for all $\mathcal{S} \in \mathscr{X}^n$; this condition might be relaxed to account for kernels with vanishing diagonal, but one might then need to introduce ad hoc conditions to ensure that $\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2$ remains large enough during the minimisation process.

3 Stochastic Approximation of the Radial-SKD Gradient

The complexity of evaluating a partial derivative of $R : \mathscr{X}^n \to \mathbb{R}$ is $\mathcal{O}(n^2 + nN)$, which might become prohibitive for large values of N. To overcome this limitation, stochastic approximations of the gradient of R might be considered (see e.g. [2]).

The evaluation of (5) involves, for instance, terms of the form $\sum_{i=1}^{N} K^2(s, x_i)$, with $s \in \mathscr{X}$ and $\mathcal{D} = \{x_1, \cdots, x_N\}$. Introducing a random variable X with uniform distribution on $\mathcal{D}$, we can note that

$$\sum_{i=1}^{N} K^2(s, x_i) = N\mathbb{E}\big[K^2(s, X)\big],$$

and the mean $\mathbb{E}[K^2(s, X)]$ may then, classically, be approximated by random sampling. More precisely, if $X_1, \cdots, X_b$ are $b \in \mathbb{N}$ copies of X, we have

$$\mathbb{E}\big[K^2(s, X)\big] = \frac{1}{b}\sum_{j=1}^{b} \mathbb{E}\big[K^2(s, X_j)\big] \text{ and } \mathbb{E}\big[\partial^{[\mathrm{l}]}_{[s]_l} K^2(s, X)\big] = \frac{1}{b}\sum_{j=1}^{b} \mathbb{E}\big[\partial^{[\mathrm{l}]}_{[s]_l} K^2(s, X_j)\big],$$

so that we can easily define unbiased estimators of the various terms appearing in (5). We refer to the sample size b as the *batch size*.

Let $k \in \{1, \ldots, n\}$ and $l \in \{1, \ldots, d\}$; the partial derivative (5) can be rewritten as

$$\partial_{[s_k]_l} R(\mathcal{S}) = \frac{T_1^2}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^4} \Upsilon(\mathcal{S}) - \frac{2T_1 T_2^{k,l}}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2},$$

with $T_1 = \sum_{i=1}^{N}\sum_{j=1}^{n} K^2(s_j, x_i)$ and $T_2^{k,l} = \sum_{i=1}^{N} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, x_i)$, and

$$\Upsilon(\mathcal{S}) = \partial^{[\mathrm{d}]}_{[s_k]_l} K^2(s_k, s_k) + 2\sum_{\substack{j=1,\\ j\neq k}}^{n} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, s_j).$$

The terms T_1 and $T_2^{k,l}$ are the only terms in (5) that depend on $\mathcal{D}$. From a random sample $\mathbf{X} = \{X_1, \cdots, X_b\}$, we define the unbiased estimators $\hat{T}_1(\mathbf{X})$ of T_1, and $\hat{T}_2^{k,l}(\mathbf{X})$ of $T_2^{k,l}$, as

$$\hat{T}_1(\mathbf{X}) = \frac{N}{b}\sum_{i=1}^{n}\sum_{j=1}^{b} K^2(s_i, X_j), \quad \text{and} \quad \hat{T}_2^{k,l}(\mathbf{X}) = \frac{N}{b}\sum_{j=1}^{b} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, X_j).$$

In what follows, we discuss the properties of some stochastic approximations of the gradient of R that can be defined from such estimators.

One-Sample Approximation. Using a single random sample $\mathbf{X}=\{X_1, \cdots, X_b\}$ of size b, we can define the following stochastic approximation of the partial derivative (5):

$$\hat{\partial}_{[s_k]_l} R(\mathcal{S}; \mathbf{X}) = \frac{\hat{T}_1(\mathbf{X})^2}{\|\mathbf{K}_\mathcal{S}\|_\mathrm{F}^4}\Upsilon(\mathcal{S}) - \frac{2\hat{T}_1(\mathbf{X})\hat{T}_2^{k,l}(\mathbf{X})}{\|\mathbf{K}_\mathcal{S}\|_\mathrm{F}^2}. \tag{6}$$

An evaluation of $\hat{\partial}_{[s_k]_l} R(\mathcal{S}; \mathbf{X})$ has complexity $\mathcal{O}(n^2 + nb)$, as opposed to $\mathcal{O}(n^2 + nN)$ for the corresponding exact partial derivative. However, due to the dependence between $\hat{T}_1(\mathbf{X})$ and $\hat{T}_2^{k,l}(\mathbf{X})$, and to the fact that $\hat{\partial}_{[s_k]_l} R(\mathcal{S}; \mathbf{X})$ involves the square of $\hat{T}_1(\mathbf{X})$, the stochastic partial derivative $\hat{\partial}_{[s_k]_l} R(\mathcal{S}; \mathbf{X})$ will generally be a biased estimator of $\partial_{[s_k]_l} R(\mathcal{S})$.

Two-Sample Approximation. To obtain an unbiased estimator of the partial derivative (5), instead of considering a single random sample, we may define a stochastic approximation based on two independent random samples $\mathbf{X} = \{X_1, \cdots, X_{b_\mathbf{X}}\}$ and $\mathbf{Y} = \{Y_1, \cdots, Y_{b_\mathbf{Y}}\}$, consisting of $b_\mathbf{X}$ and $b_\mathbf{Y} \in \mathbb{N}$ copies of X (i.e. consisting of uniform random variables on $\mathcal{D}$), with $b = b_\mathbf{X} + b_\mathbf{Y}$. The two-sample estimator of (5) is then given by

$$\hat{\partial}_{[s_k]_l} R(\mathcal{S}; \mathbf{X}, \mathbf{Y}) = \frac{\hat{T}_1(\mathbf{X})\hat{T}_1(\mathbf{Y})}{\|\mathbf{K}_\mathcal{S}\|_\mathrm{F}^4}\Upsilon(\mathcal{S}) - \frac{2\hat{T}_1(\mathbf{X})\hat{T}_2^{k,l}(\mathbf{Y})}{\|\mathbf{K}_\mathcal{S}\|_\mathrm{F}^2}, \tag{7}$$

and since $\mathbb{E}\big[\hat{T}_1(\mathbf{X})\hat{T}_1(\mathbf{Y})\big] = T_1^2$ and $\mathbb{E}\big[\hat{T}_1(\mathbf{X})\hat{T}_2^{k,l}(\mathbf{Y})\big] = T_1 T_2^{k,l}$, we have

$$\mathbb{E}\Big[\hat{\partial}_{[s_k]_l} R(\mathcal{S}; \mathbf{X}, \mathbf{Y})\Big] = \partial_{[s_k]_l} R(\mathcal{S}).$$

Although being unbiased, for a common batch size b, the variance of the two-sample estimator (7) will generally be larger than the variance of the one-sample estimator (6). In our numerical experiments, the larger variance of the unbiased estimator (7) seems to actually slow down the descent when compared to the descent obtained with the one-sample estimator (6).

Remark 3. While considering two independent samples $\mathbf{X}$ and $\mathbf{Y}$, the two terms $\hat{T}_1(\mathbf{X})\hat{T}_1(\mathbf{Y})$ and $\hat{T}_1(\mathbf{X})\hat{T}_2^{k,l}(\mathbf{Y})$ appearing in (7) are dependent. This dependence may complicate the analysis of the properties of the resulting SGD; nevertheless, this issue might be overcome by considering four independent samples instead of two. ◁

4 Numerical Experiments

Throughout this section, the matrices $\mathbf{K}$ are defined from multisets $\mathcal{D} = \{x_1, \cdots, x_N\} \subset \mathbb{R}^d$ and from kernels K of the form $K(x,t) = e^{-\rho\|x-t\|^2}$, with $\rho > 0$ and where $\|.\|$ is the Euclidean norm of $\mathbb{R}^d$ (Gaussian kernel). Except for the synthetic example of Sect. 4.1, all the multisets $\mathcal{D}$ we consider consist of the entries of data sets available on the UCI Machine Learning Repository; see [6].

Our experiments are based on the following protocol: for a given $n \in \mathbb{N}$, we consider an initial Nyström sample $\mathcal{S}^{(0)}$ consisting of n points drawn uniformly at random, without replacement, from $\mathcal{D}$. The initial sample $\mathcal{S}^{(0)}$ is regarded as an element of $\mathscr{X}^n$, and is used to initialise a SGD (except in Sect. 4.1, where GD is used), with fixed stepsize $\gamma > 0$, for the minimisation of R over $\mathscr{X}^n$, yielding, after $T \in \mathbb{N}$ iterations, a locally optimised Nyström sample $\mathcal{S}^{(T)}$. The SGDs are performed with the one-sample estimator (6) and are based on independent and identically distributed uniform random variables on $\mathcal{D}$ (i.e. i.i.d. sampling), with batch size $b \in \mathbb{N}$; see Sect. 3. We assess the accuracy of the Nyström approximations of $\mathbf{K}$ induced by $\mathcal{S}^{(0)}$ and $\mathcal{S}^{(T)}$ in terms of radial SKD and of the classical criteria (C.1)–(C.3) (for large matrices, we only consider the trace norm). We in parallel investigate the impact of the Nyström-sample size (Sects. 4.1 and 4.3) and of the kernel parameter (Sect. 4.2), and demonstrate the ability of the proposed approach to tackle problems of relatively large size (Sect. 4.4).

For a Nyström sample $\mathcal{S} \in \mathscr{X}^n$ of size $n \in \mathbb{N}$, the matrix $\hat{\mathbf{K}}(\mathcal{S})$ is of rank at most n. Following [4,10], to assess the efficiency of the approximation of $\mathbf{K}$ induced by $\mathcal{S}$, we consider the *approximation factors*

$$\mathcal{E}_{\mathrm{tr}}(\mathcal{S}) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_*}{\|\mathbf{K} - \hat{\mathbf{K}}_n^*\|_*}, \quad \mathcal{E}_{\mathrm{F}}(\mathcal{S}) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_{\mathrm{F}}}{\|\mathbf{K} - \hat{\mathbf{K}}_n^*\|_{\mathrm{F}}}, \quad \text{and } \mathcal{E}_{\mathrm{sp}}(\mathcal{S}) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_2}{\|\mathbf{K} - \hat{\mathbf{K}}_n^*\|_2}, \tag{8}$$

where $\hat{\mathbf{K}}_n^*$ denotes an optimal rank-n approximation of $\mathbf{K}$ (i.e. the approximation of $\mathbf{K}$ obtained by truncation of a spectral expansion of $\mathbf{K}$ and based on n of the largest eigenvalues of $\mathbf{K}$). The closer $\mathcal{E}_{\mathrm{tr}}(\mathcal{S})$, $\mathcal{E}_{\mathrm{F}}(\mathcal{S})$ and $\mathcal{E}_{\mathrm{sp}}(\mathcal{S})$ are to 1, the more efficient the approximation is.

4.1 Bi-Gaussian Example

We consider a kernel matrix $\mathbf{K}$ defined by a set $\mathcal{D}$ consisting of $N = 2{,}000$ points in $[-1,1]^2 \subset \mathbb{R}^2$ (i.e. $d = 2$); for the kernel parameter, we use $\rho = 1$. A graphical representation of the set $\mathcal{D}$ is given in Fig. 1; it consists of N independent realisations of a bivariate random variable whose density is proportional to the restriction of a bi-Gaussian density to the set $[-1,1]^2$ (the two modes of the underlying distribution are located at $(-0.8, 0.8)$ and $(0.8, -0.8)$, and the covariance matrix of each Gaussian density is $\mathbb{I}_2/2$, with $\mathbb{I}_2$ the 2×2 identity matrix).

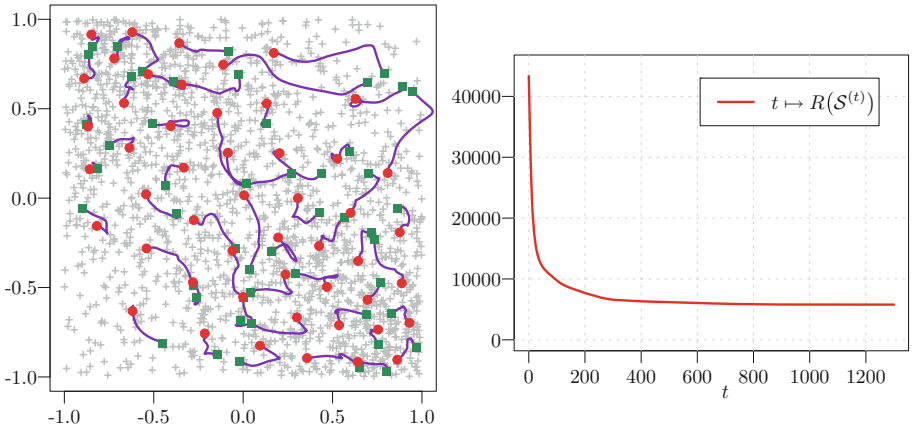

Fig. 1. For the bi-Gaussian example, graphical representation of the path $t \mapsto \mathcal{S}^{(t)}$ followed by the landmark points of a Nystrom sample during the local minimisation of R through GD, with $n = 50$, $\gamma = 10^{-6}$ and $T = 1{,}300$; the green squares are the landmark points of the initial sample $\mathcal{S}^{(0)}$, the red dots are the landmark points of the locally optimised sample $\mathcal{S}^{(T)}$, and the purple lines correspond to the paths followed by each landmark point (left). The evolution, during the GD, of the radial-SKD and trace criteria is also presented (right). (Color figure online)

The initial samples $\mathcal{S}^{(0)}$ are optimised via GD with stepsize $\gamma = 10^{-6}$ and for a fixed number of iterations T. A graphical representation of the paths followed by the landmark points during the optimisation process is given in Fig. 1 (for $n = 50$ and $T = 1{,}300$); we observe that the landmark points exhibit a relatively complex dynamic, some of them showing significant displacements from their initial positions. The optimised landmark points concentrate around the regions where the density of points in $\mathcal{D}$ is the largest, and inherit a space-filling-type property in accordance with the stationarity of the kernel K. We also observe that the minimisation of the radial-SKD criterion induces a significant decay of the trace criterion (C.1).

To assess the improvement, in terms of Nyström approximation, yielded by the optimisation of the radial-SKD, for a given number of landmark points $n \in \mathbb{N}$, we randomly draw an initial Nyström sample $\mathcal{S}^{(0)}$ from $\mathcal{D}$ (uniform sampling without replacement) and compute the corresponding locally optimised sample $\mathcal{S}^{(T)}$ (GD with $\gamma = 10^{-6}$ and $T = 1{,}000$). We then compare $R\big(\mathcal{S}^{(0)}\big)$ with $R\big(\mathcal{S}^{(T)}\big)$, and compute the corresponding approximation factors with respect to the trace, Frobenius and spectral norms, see (8). We consider three different values of n, namely $n = 20$, 50 and 80, and each time perform $m = 1{,}000$ repetitions of this experiment. Our results are presented in Fig. 2; we observe that, independently of n, the local optimisation produces a significant improvement of the Nyström approximation accuracy for all the criterion considered; the

improvements are particularly noticeable for the trace and Frobenius norms, and slightly less for the spectral norm (which of the three, appears the coarsest measure of the approximation accuracy). Remarkably, the efficiencies of the locally optimised Nyström samples are relatively close to each other, in particular in terms of trace and Frobenius norms, suggesting that a large proportion of the local minima of the radial SKD induce approximations of comparable quality.

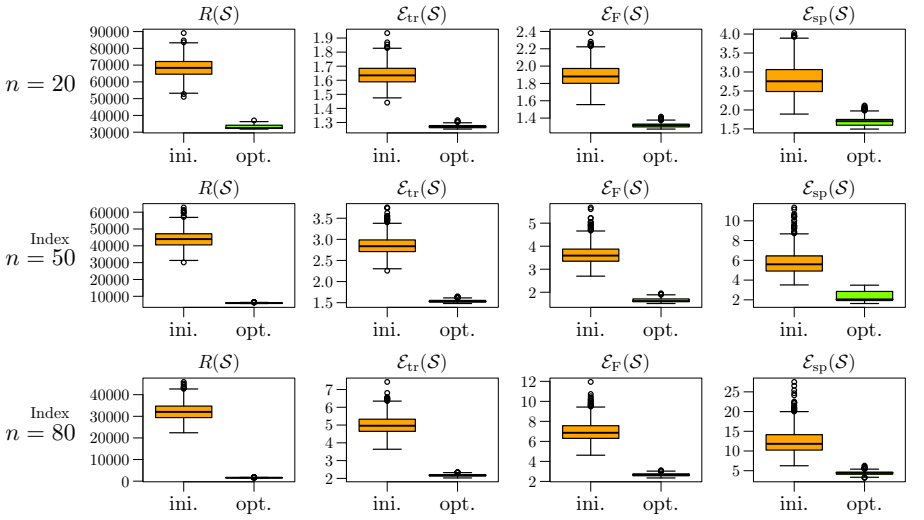

Fig. 2. For the bi-Gaussian example, comparison of the efficiency of the Nyström approximations for the initial samples $\mathcal{S}^{(0)}$ and the locally optimised samples $\mathcal{S}^{(T)}$ (optimisation through GD with $\gamma = 10^{-6}$ and $T = 1{,}000$). Each row corresponds to a given value of n; in each case $m = 1{,}000$ repetitions are performed. The first column corresponds to the radial SKD, and the following three correspond to the approximation factors defined in (8).

To further illustrate the relationship between the radial SKD and the criteria (C.1)–(C.3), for $m = 200$ random initial samples of size $n = 15$, we perform direct minimisations, through GD, of the criteria R and C_{tr} (we consider the trace norm as it is the less costly to implement). For each descent, we assess the accuracy of the locally-optimised Nyström samples in terms of radial SKD and trace norm; the results are presented in Fig. 3. We observe some strong similarities between the radial-SKD and trace-norm landscapes, further supporting the use of the radial SKD as a surrogate for the trace criterion (the minimisation of the radial SKD being, from a numerical standpoint, significantly more affordable than the minimisation of the trace norm; see Sect. 1.3).

4.2 Abalone Data Set

We now consider the $d = 8$ attributes of the Abalone data set. After removing two observations that are clear outliers, we are left with $N = 4{,}175$ entries. Each

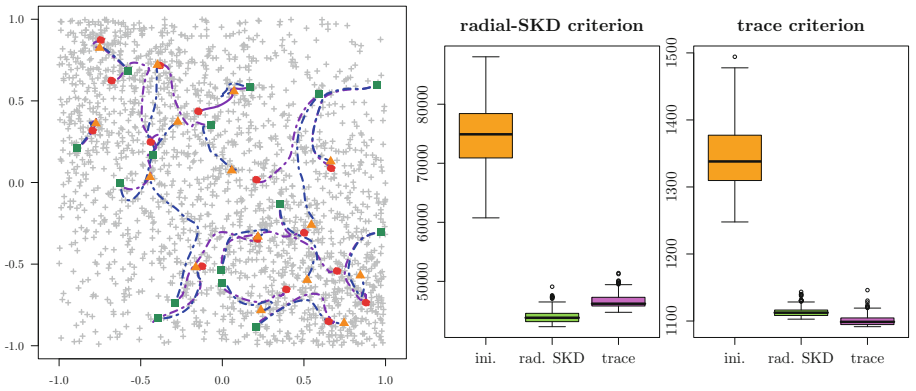

Fig. 3. For the bi-Gaussian example, graphical representation of the paths followed by the landmark points of a random initial sample of size $n = 15$ during the local minimisations of R and C_{tr} through GD; the green squares are the initial landmark points, and the red dots and orange triangles are the optimised landmark points for R and C_{tr}, respectively. The solid purple lines correspond to the paths followed by the points during the minimisation of R, and the dashed blue lines to the paths followed during the minimisation of C_{tr} (left). For $m = 200$ random initial Nyström samples of size $n = 15$, comparison of the improvements yielded by the minimisations of R and C_{tr} in terms of radial SKD (middle) and trace norm (right). Each GD uses $T = 1{,}000$ iterations, with $\gamma = 10^{-6}$ for R and $\gamma = 8 \times 10^{-5}$ for C_{tr}. (Color figure online)

of the 8 features is standardised such that it has zero mean and unit variance. We set $n = 50$ and consider three different values of the kernel paramater ρ, namely $\rho = 0.25$, 1, and 4; this values are chosen so that the eigenvalues of the kernel matrix $\mathbf{K}$ exhibit sharp, moderate and shallower decays, respectively. For the Nyström sample optimisation, we use SGD with i.i.d. sampling and batch size $b = 50$, $T = 10{,}000$ and $\gamma = 8 \times 10^{-7}$; these values were chosen to obtain relatively efficient optimisations for the whole range of values of ρ we consider. For each value of ρ, we perform $m = 200$ repetitions. The results are presented in Fig. 4.

We observe that regardless of the values of ρ and in comparison with the initial Nyström samples, the efficiencies of the locally optimised samples in terms of trace, Frobenius and spectral norms are significantly improved. As observed in Sect. 4.1, the gains yielded by the local optimisations are more evident in terms of trace and Frobenius norms, and the impact of the initialisation appears limited.

4.3 MAGIC Data Set

We consider the $d = 10$ attributes of the MAGIC Gamma Telescope data set. In pre-processing, we remove the 115 duplicated entries in the data set, leaving us with $N = 18{,}905$ data points; we then standardise each of the $d = 10$ features of the data set. For the kernel parameter, we use $\rho = 0.2$.

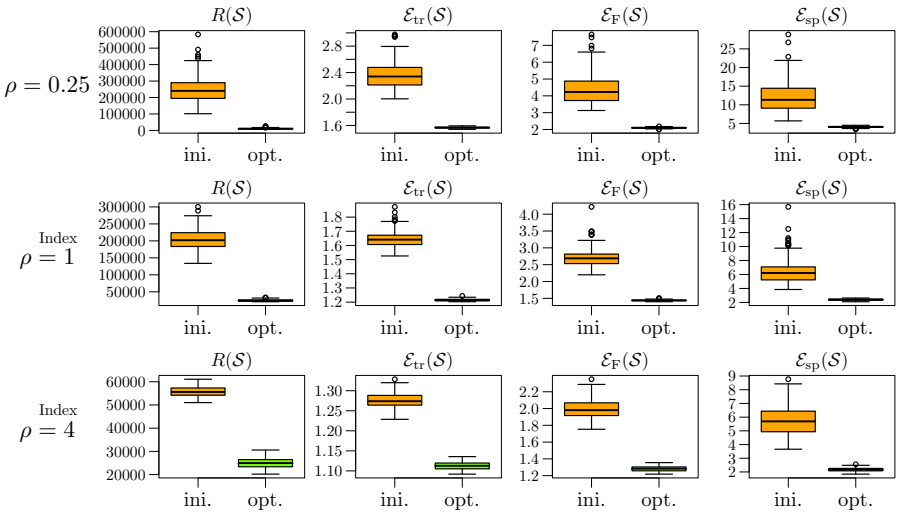

Fig. 4. For the Abalone data set with $n = 50$ and $\rho \in \{0.25, 1, 4\}$, comparison of the efficiency of the Nyström approximations for the initial Nyström samples $\mathcal{S}^{(0)}$ and the locally optimised samples $\mathcal{S}^{(T)}$ (SGD with i.i.d sampling, $b = 50$, $\gamma = 8 \times 10^{-7}$ and $T = 10{,}000$). Each row corresponds to a given value of ρ; in each case, $m = 200$ repetitions are performed.

In Fig. 5, we present the results obtained after the local optimisation of $m = 200$ random initial Nyström samples of size $n = 100$ and 200. Each optimisation was performed through SGD with i.i.d. sampling, batch size $b = 50$ and stepsize $\gamma = 5 \times 10^{-8}$; as number of iterations, for $n = 100$, we used $T = 3{,}000$, and $T = 4{,}000$ for $n = 200$. The optimisation parameters were chosen to obtain relatively efficient but not fully completed descents, as illustrated in Fig. 5. Alongside the radial SKD, we only compute the approximation factor corresponding to the trace norm (the trace norm is indeed the least costly to evaluate of the three matrix norms we consider, see Sect. 1.2). As in the previous experiments, we observe a significant improvement of the initial Nyström samples obtained by local optimisation of the radial SKD.

4.4 MiniBooNE Data Set

In this last experiment, we consider the $d = 50$ attributes of the MiniBooNE particle identification data set. In pre-processing, we remove the 471 entries in the data set with missing values, and 1 entry appearing as a clear outlier, leaving us with $N = 129{,}592$ data points; we then standardise each of the $d = 50$ features of the data set. We use $\rho = 0.04$ (kernel parameter).

We consider a random initial Nyström sample of size $n = 1{,}000$, and optimise it through SGD with i.i.d. sampling, batch size $b = 200$, stepsize $\gamma = 2 \times 10^{-7}$; the descent is stopped after $T = 8{,}000$ iterations. The resulting decay of the

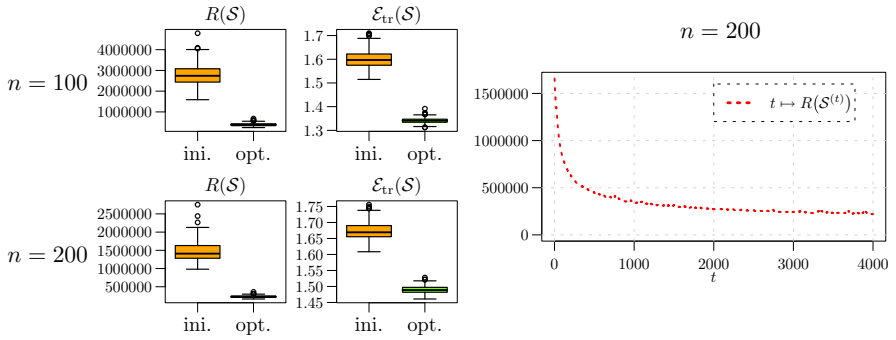

Fig. 5. For the MAGIC data set, boxplots of the radial SKD R and of the approximation factor $\mathcal{E}_{\mathrm{tr}}$ before and after the local optimisation via SGD of random Nyström samples of size $n = 100$ and 200; for each value of n, $m = 200$ repetitions are performed. The SGD is based on i.i.d. sampling, with $b = 50$ and $\gamma = 5 \times 10^{-8}$; for $n = 100$, the descent is stopped after $T = 3{,}000$ iterations, and after $T = 4{,}000$ iterations for $n = 200$ (left). A graphical representation of the decay of the radial SKD is also presented for $n = 200$ (right).

radial SKD is presented in Fig. 6 (the cost is evaluated every 100 iterations), and the trace norm of the Nyström approximation error for the initial and locally optimised samples are reported.

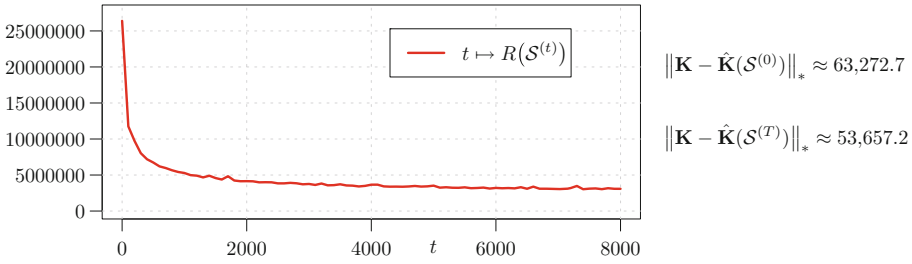

Fig. 6. For the MiniBooNE data set, decay of the radial SKD during the optimisation of a random initial Nyström sample of size $n = 1{,}000$. The SGD is based on i.i.d. sampling with batch size $b = 200$ and stepsize $\gamma = 2 \times 10^{-7}$, and the descent is stopped after $T = 8{,}000$ iterations; the cost is evaluated every 100 iterations.

In terms of computation time, on our machine (endowed with an 3.5 GHz Dual-Core Intel Core i7, and using a single-threaded C implementation interfaced with R), for $n = 1{,}000$, an evaluation of the radial SKD (up to the constant $\|\mathbf{K}\|_{\mathrm{F}}^2$) takes 6.8 s, while an evaluation of the trace criterion $\|\mathbf{K} - \hat{\mathbf{K}}(S)\|_*$ takes 6,600 s (the pseudoinverse of $\mathbf{K}_{\mathcal{S}}$ being computed in R); performing the optimisation reported in Fig. 6 without checking the decay of the cost takes 1,350 s. In

this specific setting, the full radial-SKD optimisation process is thus roughly 5 times faster than a single evaluation of the trace criterion.

5 Conclusion

We demonstrated the relevance of the radial-SKD framework for the local optimisation, through SGD, of Nyström samples for SPSD kernel-matrix approximation. We studied the Lipschitz continuity of the underlying gradient and discussed its stochastic approximation. We performed numerical experiments illustrating that local optimisation of the radial SKD yields significant improvement of the Nyström approximation in terms of trace, Frobenius and spectral norms.

In our experiments, we used SGD with i.i.d. sampling, fixed stepsize and fixed number of iterations. Although already bringing satisfactory results, to improve the efficiency of the approach, the optimisation could be accelerated by considering for instance adaptive stepsizes or momentum-type techniques (see [16] for an overview), and parallelisation may be implemented. The initial Nyström samples $\mathcal{S}^{(0)}$ we considered were drawn uniformly at random without replacement; while our experiments suggest that the local minima of the radial SKD often induce approximations of comparable quality, the use of more efficient initialisation strategies may be investigated (see e.g. [3,4,11,13,18]). To evaluate the involved partial derivatives, we relied on analytical expressions of the partial derivatives of the kernel; nevertheless, in cases where implementing such analytical expressions might prove challenging, and at the cost of a loss in computational efficiency, numerical approximation of the partial derivatives (through finite differences for instance) may be considered.

As a side note, when considering the trace norm, the Nyström sampling problem is intrinsically related to the *integrated-mean-squared-error* design criterion in kernel regression (see e.g. [8,15,17]); consequently the approach considered in this paper may be used for the design of experiments for such models.

Acknowledgments. M. Hutchings thankfully acknowledges funding from the Engineering and Physical Sciences Research Council grant EP/T517951/1. All data supporting this study is openly available in the UCI Machine Learning repository at https://archive.ics.uci.edu/.

Appendix

Proof of Theorem 1. We consider a Nyström sample $\mathcal{S} \in \mathscr{X}^n$ and introduce

$$c_{\mathcal{S}} = \frac{1}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2} \sum_{i=1}^{N} \sum_{j=1}^{n} K^2(x_i, s_j). \tag{9}$$

In view of (5), the partial derivative of R at $\mathcal{S}$ with respect to the l-th coordinate of the k-th landmark point s_k can be written as

$$\partial_{[s_k]_l} R(\mathcal{S}) = c_{\mathcal{S}}^2 \left(\partial^{[\mathrm{d}]}_{[s_k]_l} K^2(s_k, s_k) + 2 \sum_{\substack{j=1,\\ j\neq k}}^{n} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, s_j) \right) - 2c_{\mathcal{S}} \sum_{i=1}^{N} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, x_i). \tag{10}$$

For k and $k' \in \{1, \cdots, n\}$ with $k \neq k'$, and for l and $l' \in \{1, \cdots, d\}$, the second-order partial derivatives of R at $\mathcal{S}$, with respect to the coordinates of the landmark points in $\mathcal{S}$, verify

$$\begin{aligned} \partial_{[s_k]_l} \partial_{[s_k]_{l'}} R(\mathcal{S}) = {} & c_{\mathcal{S}}^2 \partial^{[\mathrm{d}]}_{[s_k]_l} \partial^{[\mathrm{d}]}_{[s_k]_{l'}} K^2(s_k, s_k) + 2c_{\mathcal{S}} (\partial_{[s_k]_{l'}} c_{\mathcal{S}}) \partial^{[\mathrm{d}]}_{[s_k]_l} K^2(s_k, s_k) \\ & + 2c_{\mathcal{S}}^2 \sum_{\substack{j=1,\\ j\neq k}}^{n} \partial^{[\mathrm{l}]}_{[s_k]_l} \partial^{[\mathrm{l}]}_{[s_k]_{l'}} K^2(s_k, s_j) + 4c_{\mathcal{S}} (\partial_{[s_k]_{l'}} c_{\mathcal{S}}) \sum_{\substack{j=1,\\ j\neq k}}^{n} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, s_j) \\ & - 2c_{\mathcal{S}} \sum_{i=1}^{N} \partial^{[\mathrm{l}]}_{[s_k]_l} \partial^{[\mathrm{l}]}_{[s_k]_{l'}} K^2(s_k, x_i) - 2(\partial_{[s_k]_{l'}} c_{\mathcal{S}}) \sum_{i=1}^{N} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, x_i), \text{ and} \end{aligned} \tag{11}$$

$$\begin{aligned} \partial_{[s_k]_l} \partial_{[s_{k'}]_{l'}} R(\mathcal{S}) = {} & 2c_{\mathcal{S}} (\partial_{[s_{k'}]_{l'}} c_{\mathcal{S}}) \partial^{[\mathrm{d}]}_{[s_k]_l} K^2(s_k, s_k) + 2c_{\mathcal{S}}^2 \partial^{[\mathrm{l}]}_{[s_k]_l} \partial^{[\mathrm{r}]}_{[s_{k'}]_{l'}} K^2(s_k, s_{k'}) \\ & + 4c_{\mathcal{S}} (\partial_{[s_{k'}]_{l'}} c_{\mathcal{S}}) \sum_{\substack{j=1,\\ j\neq k}}^{n} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, s_j) - 2(\partial_{[s_{k'}]_{l'}} c_{\mathcal{S}}) \sum_{i=1}^{N} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, x_i), \end{aligned} \tag{12}$$

the partial derivative of $c_{\mathcal{S}}$ with respect to the l-th coordinate of the k-th landmark point s_k is given by

$$\partial_{[s_k]_l} c_{\mathcal{S}} = \frac{1}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2} \left(\sum_{i=1}^{N} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, x_i) - c_{\mathcal{S}} \partial^{[\mathrm{d}]}_{[s_k]_l} K^2(s_k, s_k) - 2c_{\mathcal{S}} \sum_{\substack{j=1,\\ j\neq k}}^{n} \partial^{[\mathrm{l}]}_{[s_k]_l} K^2(s_k, s_j) \right). \tag{13}$$

From (A.1), we have

$$\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} K^2(s_i, s_j) \geqslant \sum_{i=1}^{n} K^2(s_i, s_i) \geqslant n\alpha. \tag{14}$$

By the Schur product theorem, the squared kernel K^2 is SPSD; we denote by $\mathcal{G}$ the RKHS of real-valued functions on $\mathscr{X}$ for which K^2 is reproducing. For x and $y \in \mathscr{X}$, we have $K^2(x,y) = \langle k_x^2, k_y^2 \rangle_{\mathcal{G}}$, with $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ the inner product on $\mathcal{G}$, and where $k_x^2 \in \mathcal{G}$ is such that $k_x^2(t) = K^2(t,x)$, for all $t \in \mathscr{X}$. From the Cauchy-Schwartz inequality, we have

$$\sum_{i=1}^{N}\sum_{j=1}^{n} K^2(s_j, x_i) = \sum_{i=1}^{N}\sum_{j=1}^{n} \langle k_{s_j}^2, k_{x_i}^2 \rangle_{\mathcal{G}} = \left\langle \sum_{j=1}^{n} k_{s_j}^2, \sum_{i=1}^{N} k_{x_i}^2 \right\rangle_{\mathcal{G}}$$
$$\leqslant \left\| \sum_{j=1}^{n} k_{s_j}^2 \right\|_{\mathcal{G}} \left\| \sum_{i=1}^{N} k_{x_i}^2 \right\|_{\mathcal{G}} = \|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}} \|\mathbf{K}\|_{\mathrm{F}}. \tag{15}$$

By combining (9) with inequalities (14) and (15), we obtain

$$0 \leqslant c_{\mathcal{S}} \leqslant \frac{\|\mathbf{K}\|_{\mathrm{F}}}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}} \leqslant \frac{\|\mathbf{K}\|_{\mathrm{F}}}{\sqrt{n\alpha}} = C_0. \tag{16}$$

Let $k \in \{1, \ldots, n\}$ and let $l \in \{1, \ldots, d\}$; from Eq. (13), and using inequalities (14) and (16) together with (A.2), we obtain

$$|\partial_{[s_k]_l} c_{\mathcal{S}}| \leqslant \frac{M_1}{n\alpha}[N + (2n-1)C_0] = C_1. \tag{17}$$

In addition, let $k' \in \{1, \ldots, n\} \setminus \{k\}$ and $l' \in \{1, \ldots, d\}$; from Eqs. (11), (12), (16) and (17), and conditions (A.2) and (A.3), we get

$$\begin{aligned} &|\partial_{[s_k]_l} \partial_{[s_k]_{l'}} R(\mathcal{S})| \\ &\quad \leqslant C_0^2 M_2 + 2C_0 C_1 M_1 + 2(n-1)C_0^2 M_2 + 4(n-1)C_0 C_1 M_1 + 2C_0 M_2 N + 2C_1 M_1 N \\ &\quad = (2n-1)C_0^2 M_2 + (4n-2)C_0 C_1 M_1 + 2N(C_0 M_2 + C_1 M_1), \end{aligned} \tag{18}$$

and

$$\begin{aligned} |\partial_{[s_k]_l} \partial_{[s_{k'}]_{l'}} R(\mathcal{S})| &\leqslant 2C_0 C_1 M_1 + 2C_0^2 M_2 + 4(n-1)C_0 C_1 M_1 + 2C_1 M_1 N \\ &= 2C_0^2 M_2 + (4n-2)C_0 C_1 M_1 + 2\,N C_1 M_1. \end{aligned} \tag{19}$$

For $k, k' \in \{1, \ldots, n\}$, we denote by $\mathbf{B}^{k,k'}$ the $d \times d$ matrix with l, l' entry given by (11) if $k = k'$, and by (12) otherwise. The Hessian $\nabla^2 R(\mathcal{S})$ can then be represented as a block-matrix, that is

$$\nabla^2 R(\mathcal{S}) = \begin{bmatrix} \mathbf{B}^{1,1} & \cdots & \mathbf{B}^{1,n} \\ \vdots & \ddots & \vdots \\ \mathbf{B}^{n,1} & \cdots & \mathbf{B}^{n,n} \end{bmatrix} \in \mathbb{R}^{nd \times nd}.$$

The d^2 entries of the n diagonal blocks of $\nabla^2 R(\mathcal{S})$ are of the form (11), and the d^2 entries of the $n(n-1)$ off-diagonal blocks of $\nabla^2 R(\mathcal{S})$ are the form (12). From inequalities (18) and (19), we obtain

$$\|\nabla^2 R(\mathcal{S})\|_2^2 \leqslant \|\nabla^2 R(\mathcal{S})\|_{\mathrm{F}}^2 = \sum_{k=1}^{n}\sum_{l=1}^{d}\sum_{l'=1}^{d}[\mathbf{B}^{k,k}]_{l,l'}^2 + \sum_{k=1}^{n}\sum_{\substack{k'=1,\\ k'\neq k}}^{n}\sum_{l=1}^{d}\sum_{l'=1}^{d}[\mathbf{B}^{k,k'}]_{l,l'}^2 \leqslant L^2,$$

with

$$\begin{aligned} L = \big(&nd^2[(2n-1)C_0^2M_2 + (4n-2)C_0C_1M_1 + 2N(C_0M_2 + C_1M_1)]^2 \\ &+ 4n(n-1)d^2[C_0^2M_2 + (2n-1)C_0C_1M_1 + NC_1M_1]^2\big)^{\frac{1}{2}}. \end{aligned}$$

For all $\mathcal{S} \in \mathscr{X}^n$, the constant L is an upper bound for the spectral norm of the Hessian matrix $\nabla^2 R(\mathcal{S})$, so the gradient of R is Lipschitz continuous over $\mathscr{X}^n$, with Lipschitz constant L. □

References

1. Berlinet, A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer, New York (2004). https://doi.org/10.1007/978-1-4419-9096-9
2. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018)
3. Cai, D., Chow, E., Erlandson, L., Saad, Y., Xi, Y.: SMASH: structured matrix approximation by separation and hierarchy. Numer. Linear Algebra Appl. **25** (2018)
4. Derezinski, M., Khanna, R., Mahoney, M.W.: Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nyström method. In: Advances in Neural Information Processing Systems (2020)
5. Drineas, P., Mahoney, M.W.: On the Nyström method for approximating a Gram matrix for improved kernel-based learning. J. Mach. Learn. Res. **6**, 2153–2175 (2005)
6. Dua, D., Graff, C.: UCI Machine Learning Repository (2019). http://archive.ics.uci.edu/ml
7. Gauthier, B.: Nyström approximation and reproducing kernels: embeddings, projections and squared-kernel discrepancy. Preprint (2021). https://hal.archives-ouvertes.fr/hal-03207443
8. Gauthier, B., Pronzato, L.: Convex relaxation for IMSE optimal design in random-field models. Comput. Stat. Data Anal. **113**, 375–394 (2017)
9. Gauthier, B., Suykens, J.: Optimal quadrature-sparsification for integral operator approximation. SIAM J. Sci. Comput. **40**, A3636–A3674 (2018)
10. Gittens, A., Mahoney, M.W.: Revisiting the Nyström method for improved large-scale machine learning. J. Mach. Learn. Res. **17**, 1–65 (2016)
11. Kumar, S., Mohri, M., Talwalkar, A.: Sampling methods for the Nyström method. J. Mach. Learn. Res. **13**, 981–1006 (2012)
12. Lee, J.D., Simchowitz, M., Jordan, M.I., Recht, B.: Gradient descent only converges to minimizers. In: Conference on Learning Theory, pp. 1246–1257. PMLR (2016)

13. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM (1992)
14. Paulsen, V.I., Raghupathi, M.: An Introduction to the Theory of Reproducing Kernel Hilbert Spaces. Cambridge University Press, Cambridge (2016)
15. Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
16. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
17. Santner, T.J., Williams, B.J., Notz, W.I.: The Design and Analysis of Computer Experiments. Springer, New York (2018). https://doi.org/10.1007/978-1-4757-3799-8
18. Wang, S., Zhang, Z., Zhang, T.: Towards more efficient SPSD matrix approximation and CUR matrix decomposition. J. Mach. Learn. Res. **17**, 7329–7377 (2016)