Energy-based sequential sampling for low-rank PSD-matrix approximation

Matthew Hutchings, Bertrand Gauthier

▶ **To cite this version:**

Matthew Hutchings, Bertrand Gauthier. Energy-based sequential sampling for low-rank PSD-matrix approximation. 2023. hal-04102664v1

HAL Id: hal-04102664

https://hal.science/hal-04102664v1

Preprint submitted on 22 May 2023 (v1), last revised 18 Dec 2023 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Energy-based sequential sampling for low-rank PSD-matrix approximation

Matthew Hutchings
Cardiff University, School of Mathematics
Abacws, Senghennydd Road
Cardiff, CF24 4AG, United Kingdom
HutchingsM1@cardiff.ac.uk

Bertrand Gauthier
Cardiff University, School of Mathematics
Abacws, Senghennydd Road
Cardiff, CF24 4AG, United Kingdom
GauthierB@cardiff.ac.uk

Abstract

We introduce a pseudoconvex differentiable relaxation of the column-sampling problem for the Nyström approximation of positive-semidefinite (PSD) matrices. The relaxation is based on the interpretation of PSD matrices as integral operators, and uses the supports of measures to characterise samples of columns. We describe a class of gradient-based sequential sampling strategies which leverages the properties of the considered framework and demonstrate its ability to produce accurate Nyström approximations. As an important feature, the proposed strategies rely on an isometric representation of weighted PSD matrices as potentials to efficiently handle matrices of very large scale.

Keywords: Nyström approximation, reproducing kernel Hilbert spaces, kernel quadrature, generalised convexity, conditional gradient.

1 Introduction

The low-rank approximation of matrices through column sampling is a core technique in machine learning and scientific computing. Such approximations provide a computationaly efficient framework to reduce the cost of numerical strategies involving large-scale matrices, and the underlying sampling problem is intrinsically related to feature extraction, clustering and dimensionality reduction. For positive-semidefinite (PSD) matrices, the terminology *Nyström approximation* is used, and the characterisation of samples of columns leading to accurate approximations is referred to as the *column sampling problem* (CSP); see e.g. [24, 19, 1, 23, 22]. In practical applications, the combinatorial nature of the CSP and the cost inherent to the evaluation of the approximation errors prevent the implementation of sampling strategies based on direct minimisations, and as such, have motivated the development of a wide variety of heuristic-based sampling strategies; see [5, 11, 9, 17, 4] and references therein for an overview.

In this note, we characterise samples of columns by the nonnull entries of *selection vectors* (interestingly, this alone leads to a convex, but nondifferentiable, relaxation of the CSP; see Theorem 2.1); such selection vectors can be regarded as measures, and together with a PSD matrix, define integral operators acting on the reproducing kernel Hilbert space (RKHS; see e.g. [16]) defined by this matrix. Following [8, 7], we use the norm of the corresponding Hilbert-Schmidt (HS) space to discriminate among selection vectors. Enforcing an invariance with respect to rescaling gives rise to a pseudoconvex differentiable error map R on the selection-vector space, and the gradient of this map can then be used to characterise samples of columns. We illustrate the ability of the proposed approach to generate accurate Nyström approximations and to efficiently handle very large PSD matrices.

The main contributions of this work are the study of the properties error map R (Theorem 2.2), and the introduction of a general framework allowing for the implementation of accurate and numerically efficient line-search-based sequential sampling strategies for Nyström approximation (Algorithm 1 and its variants). Sketches of the proofs of the main theoretical results are included in the body of paper, and detailed versions of the proofs of all the presented results are provided in appendix, together with additional numerical experiments.

2 Overall framework and notations

Throughout this work, we use the classical *matrix notation* and identify a vector of $\boldsymbol{\alpha} \in \mathbb{C}^N$, $N \in \mathbb{N}$, as the $N \times 1$ column matrix defined by the coefficients of $\boldsymbol{\alpha}$ in the canonical basis of $\mathbb{C}^N$. We denote by $[N]$ the set of all integers between 1 and N. For generality, we consider complex PSD matrices; nevertheless, all the developments presented in this note also hold for real symmetric positive-semidefinite (SPSD) matrices. We denote by $\mathbf{M}^*$ the conjugate-transpose of a matrix $\mathbf{M}$, and by $\overline{\mathbf{M}}$ its conjugate.

2.1 Nyström approximation of PSD matrices

Let $\mathbf{K} \in \mathbb{C}^{N\times N}$ be a PSD matrix, with $N \in \mathbb{N}$; we assume that all the diagonal entries of $\mathbf{K}$ are strictly positive (otherwise, the corresponding rows and columns of $\mathbf{K}$ are null). For a subset $I \subseteq [N]$ of size $m \leqslant N$, the *Nyström approximation* of $\mathbf{K}$ induced by the columns of $\mathbf{K}$ with index in I is the PSD matrix

$$\hat{\mathbf{K}}(I) = \mathbf{K}_{\bullet,I}(\mathbf{K}_{I,I})^{\dagger}\mathbf{K}_{I,\bullet} \in \mathbb{C}^{N\times N}, \tag{1}$$

where $\mathbf{K}_{\bullet,I} \in \mathbb{C}^{N\times m}$ is the matrix defined by the columns of $\mathbf{K}$ with index in I, and where $(\mathbf{K}_{I,I})^{\dagger}$ is the pseudoinverse of the $m \times m$ principal submatrix of $\mathbf{K}$ defined by I (and $\mathbf{K}_{I,\bullet} = (\mathbf{K}_{\bullet,I})^*$ consists of rows of $\mathbf{K}$); see e.g. [5, 18, 11, 9, 3]

The accuracy of a Nyström approximation is often assessed through the trace, Frobenius or spectral norm of the approximation error, that is

$$\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{tr}}, \quad \|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{F}}, \quad \text{or} \quad \|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{sp}}, \tag{2}$$

respectively, naturally raising questions related to the characterisation of subsets I leading to accurate approximations. In practice, a direct minimisation, as functions of I, of the error norms (2) is made difficult by the combinatorial nature of the underlying problems, and by the numerical cost inherent to the evaluation of the corresponding norms.

Remark 2.1. In view of (1), for a subset I of size m, the evaluation of a given entry of $\hat{\mathbf{K}}(I)$ requires the pseudoinversion of $K_{I,I}$, with complexity $\mathcal{O}(m^3)$, combined with an algebraic operation of complexity $\mathcal{O}(m^2)$. In the case of multiple evaluations, the pseudoinversion can be mutualised, so that in (2), the complexity of an evaluation of the trace norm is $\mathcal{O}(m^3 + m^2N)$; it is $\mathcal{O}(m^3 + m^2N^2)$ for the Frobenius norm, and $\mathcal{O}(m^3 + m^2N^2 + N^3)$ for the spectral norm. As $[\hat{\mathbf{K}}(I)]_{\bullet,I} = \mathbf{K}_{\bullet,I}$, it is in practice only necessary to compute the required entries of the $(N-m)\times(N-m)$ principal submatrix $[\hat{\mathbf{K}}(I)]_{I^c,I^c}$, with I^c the complement of I in $[N]$. ◁

Remark 2.2. The entries of $\mathbf{K}$ characterise the kernel of a RKHS of $\mathbb{C}$-valued functions on $[N]$ (see e.g. [16]). This RKHS can be identified with the subspace $\mathcal{H} = \mathrm{span}\{\mathbf{K}\} \subseteq \mathbb{C}^N$ (the subset of $\mathbb{C}^N$ spanned by the columns of $\mathbf{K}$) endowed with the inner product $\langle \boldsymbol{h} \,|\, \boldsymbol{f}\rangle_{\mathcal{H}} = \boldsymbol{h}^*\mathbf{K}^{\dagger}\boldsymbol{f}$, $\boldsymbol{h}$ and $\boldsymbol{f} \in \mathcal{H}$. A subset $I \subseteq [N]$ then defines a closed linear subspace $\mathcal{H}_I = \mathrm{span}\{\mathbf{K}_{\bullet,I}\}$ of $\mathcal{H}$, and $P_I = \mathbf{K}_{\bullet,I}(\mathbf{K}_{I,I})^{\dagger}\mathbb{I}_{I,\bullet}$ is the orthogonal projection from $\mathcal{H}$ onto $\mathcal{H}_I$, with $\mathbb{I}$ the $N \times N$ identity matrix. We then in particular have $\hat{\mathbf{K}}(I) = P_I\mathbf{K}$.

Denoting by $\mathcal{E}$ the Euclidean Hilbert space $\mathbb{C}^N$ (that is $\langle \mathbf{u} \,|\, \mathbf{v}\rangle_{\mathcal{E}} = \mathbf{u}^*\mathbf{v}$, $\mathbf{u}$ and $\mathbf{v} \in \mathcal{E}$), the matrix $\mathbf{K}$ may be regarded as an operator from, and to, $\mathcal{E}$ or $\mathcal{H}$ (four possibilities). In (2), the trace norm then corresponds to the squared HS norm of the matrix $\mathbf{K} - \hat{\mathbf{K}}(I)$ when interpreted as an operator from $\mathcal{E}$ to $\mathcal{H}$; and the Frobenius and spectral norms correspond to the HS and spectral norms, respectively, of the matrix $\mathbf{K} - \hat{\mathbf{K}}(I)$ when interpreted as an operator on $\mathcal{E}$. See [7] for a further discussion. ◁

2.2 First relaxation: selection vectors

For $\boldsymbol{\upsilon} = (\upsilon_i)_{i\in[N]} \in \mathbb{R}^N$, we set $I_{\boldsymbol{\upsilon}} = \{i \in [N] | \upsilon_i \neq 0\}$ and we refer to $I_{\boldsymbol{\upsilon}}$ as the *support* of $\boldsymbol{\upsilon}$ (we may more generally assume that $\boldsymbol{\upsilon}$ is complex, but this is not useful in the framework of this note). Through its support, a *selection vector* $\boldsymbol{\upsilon}$ naturally characterises a subset of columns of $\mathbf{K}$. Following Remark 2.2, we introduce the following simplified notations

$$\hat{\mathbf{K}}(\boldsymbol{\upsilon}) = \hat{\mathbf{K}}(I_{\boldsymbol{\upsilon}}), \quad \mathcal{H}_{\boldsymbol{\upsilon}} = \mathcal{H}_{I_{\boldsymbol{\upsilon}}} \quad \text{and} \quad P_{\boldsymbol{\upsilon}} = P_{I_{\boldsymbol{\upsilon}}}.$$

We then define the error maps

$$C_{\mathrm{tr}} : \boldsymbol{\upsilon} \mapsto \|\mathbf{K} - \hat{\mathbf{K}}(\boldsymbol{\upsilon})\|_{\mathrm{tr}}, \quad C_{\mathrm{F}} : \boldsymbol{\upsilon} \mapsto \|\mathbf{K} - \hat{\mathbf{K}}(\boldsymbol{\upsilon})\|_{\mathrm{F}}^2 \quad \text{and} \quad C_{\mathrm{sp}} : \boldsymbol{\upsilon} \mapsto \|\mathbf{K} - \hat{\mathbf{K}}(\boldsymbol{\upsilon})\|_{\mathrm{sp}}^2;$$

notice that we consider the squared Frobenius and spectral norms in the definition of C_{F} and C_{sp}.

Theorem 2.1. *The maps C_{tr}, C_{F} and C_{sp} are convex on the convex cone $\mathbb{R}^N_{\geqslant 0}$, and their directional derivatives take values in the discrete set $\{-\infty, 0\}$.*

Sketch of proof. For $J \subseteq I \subseteq [N]$, we have $\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{X}} \leqslant \|\mathbf{K} - \hat{\mathbf{K}}(J)\|_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$. Also, for $\boldsymbol{\xi} = \boldsymbol{\upsilon} + \rho(\boldsymbol{\eta} - \boldsymbol{\upsilon})$, $\boldsymbol{\upsilon}, \boldsymbol{\eta} \in \mathbb{R}^N_{\geqslant 0}$, $\rho \in (0, 1)$, we have $I_{\boldsymbol{\xi}} = I_{\boldsymbol{\upsilon}} \cup I_{\boldsymbol{\eta}}$. The map $\rho \mapsto C_{\mathrm{X}}(\boldsymbol{\upsilon} + \rho[\boldsymbol{\eta} - \boldsymbol{\upsilon}])$ is thus constant on the open interval $(0, 1)$, and we have $C_{\mathrm{X}}(\boldsymbol{\xi}) \leqslant C_{\mathrm{X}}(\boldsymbol{\upsilon})$ and $C_{\mathrm{X}}(\boldsymbol{\xi}) \leqslant C_{\mathrm{X}}(\boldsymbol{\eta})$. □

Theorem 2.1 illustrates that the error maps induced by (2) can be regarded as convex piecewise-constant maps on $\mathbb{R}^N_{\geqslant 0}$; see Figure 1 for an illustration. The selection-vector setting may to this extent be regarded as a *nondifferentiable convex relaxation* of the CSP.

2.3 Second relaxation: kernel-quadrature setting

A selection vector $\boldsymbol{\upsilon} \in \mathbb{R}^N$ can be regarded as a signed measure on $[N]$, and as such, defines together with $\mathbf{K}$ an integral operator of the form $\mathbf{KV}$, with $\mathbf{V} = \mathrm{diag}(\boldsymbol{\upsilon}) \in \mathbb{C}^{N \times N}$ the diagonal matrix with diagonal $\boldsymbol{\upsilon}$. We have $\mathbf{KV} = P_{\boldsymbol{\upsilon}}\mathbf{KV} = \mathbf{KV}P_{\boldsymbol{\upsilon}} = P_{\boldsymbol{\upsilon}}\mathbf{KV}P_{\boldsymbol{\upsilon}}$, and both the matrices $\mathbf{KV}$ and $\hat{\mathbf{K}}(\boldsymbol{\upsilon})$ have range included in $\mathcal{H}_{\boldsymbol{\upsilon}}$.

Let $\boldsymbol{\omega} \in \mathbb{R}^N$ be another selection vector, and set $\mathbf{W} = \mathrm{diag}(\boldsymbol{\omega})$. Denoting by $\mathrm{HS}(\mathcal{H})$ the Hilbert space of all HS operators on $\mathcal{H}$, we have

$$\langle \mathbf{KW} \mid \mathbf{KV} \rangle_{\mathrm{HS}(\mathcal{H})} = \boldsymbol{\omega}^* \mathbf{S} \boldsymbol{\upsilon}, \tag{3}$$

where $\mathbf{S} = \overline{\mathbf{K}} \odot \mathbf{K}$ (element-wise product) is the $N \times N$ SPSD matrix with i, j entry $|\mathbf{K}_{i,j}|^2$, the squared modulus of the i, j entry of $\mathbf{K}$; the real matrix $\mathbf{S}$ is SPSD. We may in addition notice that the Nyström approximation $\hat{\mathbf{K}}(\boldsymbol{\upsilon})$ is the orthogonal projection, in $\mathrm{HS}(\mathcal{H})$, of $\mathbf{K}$ onto the closed linear subspace of all matrices with range included in $\mathcal{H}_{\boldsymbol{\upsilon}}$. See [7] for an in-depth discussion.

Remark 2.3. Following Remark 2.2, the PSD matrix $\mathbf{S}$ defines a RKHS that can be identified with the vector space $\mathcal{G} = \mathrm{span}\{\mathbf{S}\} \subseteq \mathbb{C}^N$ endowed with the inner product $\langle \boldsymbol{g} \mid \boldsymbol{j} \rangle_{\mathcal{G}} = \boldsymbol{g}^* \mathbf{S}^\dagger \boldsymbol{j}$, $\boldsymbol{g}$ and $\boldsymbol{j} \in \mathcal{G}$. In view of (3), we have

$$\langle \mathbf{KW} \mid \mathbf{KV} \rangle_{\mathrm{HS}(\mathcal{H})} = \boldsymbol{\omega}^* \mathbf{S} \boldsymbol{\upsilon} = \boldsymbol{\omega}^* \mathbf{S}\mathbf{S}^\dagger \mathbf{S} \boldsymbol{\upsilon} = \langle \mathbf{S}\boldsymbol{\omega} \mid \mathbf{S}\boldsymbol{\upsilon} \rangle_{\mathcal{G}}, \ \boldsymbol{\upsilon} \text{ and } \boldsymbol{\omega} \in \mathbb{R}^N.$$

so that the vector $\mathbf{S}\boldsymbol{\upsilon} \in \mathcal{G}$ is an *isometric representation* of the matrix $\mathbf{KV}$ when interpreted as an HS operator on $\mathcal{H}$. We refer to $\mathbf{S}\boldsymbol{\upsilon}$ as the *potential* of $\boldsymbol{\upsilon}$ in $\mathcal{G}$, and to $\|\mathbf{S}\boldsymbol{\upsilon}\|_{\mathcal{G}}^2 = \|\mathbf{KV}\|_{\mathrm{HS}(\mathcal{H})}^2 = \boldsymbol{\upsilon}^* \mathbf{S} \boldsymbol{\upsilon}$ as the *energy* of $\boldsymbol{\upsilon}$ with respect to $\mathbf{S}$. The norm $\|\mathbf{KW} - \mathbf{KV}\|_{\mathrm{HS}(\mathcal{H})}$ can then be interpreted as a generalised *integral probability metric*, or *maximum mean discrepancy* (see e.g. [20, 15]), between the signed measures on $[N]$ defined by $\boldsymbol{\omega}$ and $\boldsymbol{\upsilon} \in \mathbb{R}^N$. Introducing $\mathbb{1} = (1)_{i \in [N]} \in \mathbb{R}^N$, we in particular have $\mathrm{diag}(\mathbb{1}) = \mathbb{I}$, and $\|\mathbf{K}\|_{\mathrm{HS}(\mathcal{H})}^2 = \mathbb{1}^* \mathbf{S} \mathbb{1} = \|\mathbf{K}\|_{\mathrm{F}}^2$. ◁

Following (3) and Remark 2.3, we define the error map $D : \mathbb{R}^N \to \mathbb{R}_{\geqslant 0}$, such that

$$D(\boldsymbol{\upsilon}) = \|\mathbf{K} - \mathbf{KV}\|_{\mathrm{HS}(\mathcal{H})}^2 = (\mathbb{1} - \boldsymbol{\upsilon})^* \mathbf{S} (\mathbb{1} - \boldsymbol{\upsilon}) = \|\mathbf{K}\|_{\mathrm{F}}^2 + \boldsymbol{\upsilon}^* \mathbf{S} \boldsymbol{\upsilon} - 2\mathfrak{g}^* \boldsymbol{\upsilon}, \boldsymbol{\upsilon} \in \mathbb{R}^N, \tag{4}$$

with $\mathfrak{g} = \mathbf{S}\mathbb{1} \in \mathbb{R}^N$ the potential of $\mathbb{1}$ in $\mathcal{G}$. The map D is convex on $\mathbb{R}^N$, and the gradient of D at $\boldsymbol{\upsilon}$ is given by $\nabla D(\boldsymbol{\upsilon}) = 2\mathbf{S}(\boldsymbol{\upsilon} - \mathbb{1}) = 2(\mathbf{S}\boldsymbol{\upsilon} - \mathfrak{g})$.

Remark 2.4. The numerical complexity of the computation of the *target potential* $\mathfrak{g} = \mathbf{S}\mathbb{1} \in \mathbb{R}^N$ is $\mathcal{O}(N^2)$; this operation can nevertheless be easily parallelised. From $\mathfrak{g}$, and assuming that the support of $\boldsymbol{\upsilon}$ is of size $m \leqslant N$, the complexity of the evaluation of $\mathfrak{g}^* \boldsymbol{\upsilon}$ is $\mathcal{O}(m)$. The complexity of the evaluation of $\boldsymbol{\upsilon}^* \mathbf{S} \boldsymbol{\upsilon}$ is $\mathcal{O}(m^2)$. ◁

Recalling that $\hat{\mathbf{K}}(I) = P_I \mathbf{K}$, $I \subseteq [N]$, from Remark 2.2, we have

$$\|\mathbf{K} - P_I \mathbf{K}\|_{\mathrm{HS}(\mathcal{H})}^2 = \mathrm{trace}\left(\mathbf{K}[\mathbf{K} - \hat{\mathbf{K}}(I)]\right) \text{ and } \|\mathbf{K} - P_I \mathbf{K} P_I\|_{\mathrm{HS}(\mathcal{H})}^2 = \|\mathbf{K}\|_{\mathrm{F}}^2 - \|\hat{\mathbf{K}}(I)\|_{\mathrm{F}}^2,$$

which suggests the definition of the two additional error maps

$$C_{\mathrm{P}}(\boldsymbol{v}) = \mathrm{trace}\left(\mathbf{K}[\mathbf{K} - \hat{\mathbf{K}}(\boldsymbol{v})]\right) \quad \text{and} \quad C_{\mathrm{PP}}(\boldsymbol{v}) = \|\mathbf{K}\|_{\mathrm{F}}^2 - \|\hat{\mathbf{K}}(\boldsymbol{v})\|_{\mathrm{HS}(\mathcal{H})}^2, \boldsymbol{v} \in \mathbb{R}^N;$$

these maps are of the same nature as the maps C_{X}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$ (Corollary 2.1; see also [7]).

Corollary 2.1. *The maps C_{P} and C_{PP} are convex on the convex cone $\mathbb{R}^N_{\geqslant 0}$, and their directional derivatives take values in the discrete set $\{-\infty, 0\}$.*

2.4 Invariance under rescaling

For $\boldsymbol{v} \in \mathbb{R}^N$ and $c > 0$, we have $I_{\boldsymbol{v}} = I_{c\boldsymbol{v}}$, and the error maps C_{X}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}, \mathrm{P}, \mathrm{PP}\}$, are thus invariant under rescaling, that is, $C_{\mathrm{X}}(c\boldsymbol{v}) = C_{\mathrm{X}}(\boldsymbol{v})$. To enforce a similar invariance within the framework of (4), we introduce the error map

$$R(\boldsymbol{v}) = \min_{c \geqslant 0} D(c\boldsymbol{v}) = D(c_{\boldsymbol{v}}\boldsymbol{v}), \boldsymbol{v} \in \mathbb{R}^N,$$

with $c_{\boldsymbol{v}} = \boldsymbol{v}^*\mathbf{S}\mathbb{1}/\boldsymbol{v}^*\mathbf{S}\boldsymbol{v}$ if $\boldsymbol{v} \in \mathscr{D} = \{\boldsymbol{v} \in \mathbb{R}^N | \boldsymbol{v}^*\mathbf{S}\mathbb{1} > 0\}$, and $c_{\boldsymbol{v}} = 0$ otherwise. In particular, for $\boldsymbol{v} \in \mathscr{D}$, we have $R(\boldsymbol{v}) = \|\mathbf{K}\|_{\mathrm{F}}^2 - (\boldsymbol{v}^*\mathbf{S}\mathbb{1})^2/(\boldsymbol{v}^*\mathbf{S}\boldsymbol{v})$. For $\boldsymbol{\eta} \in \mathbb{R}^N$, the directional derivative $\Theta(\boldsymbol{v}; \boldsymbol{\eta})$ of R at $\boldsymbol{v} \in \mathbb{R}^N$ in the direction $\boldsymbol{\eta} - \boldsymbol{v}$ is

$$\Theta(\boldsymbol{v}; \boldsymbol{\eta}) = \lim_{\rho \to 0^+} \frac{1}{\rho}\big[R\big(\boldsymbol{v} + \rho(\boldsymbol{\eta} - \boldsymbol{v})\big) - R(\boldsymbol{v})\big] = \begin{cases} -\infty \text{ if } \boldsymbol{v} \in \mathscr{Z} \text{ and } \boldsymbol{\eta} \in \mathscr{D}, \\ 2c_{\boldsymbol{v}}\mathbf{S}(c_{\boldsymbol{v}}\boldsymbol{v} - \mathbb{1}) \text{ otherwise,} \end{cases} \tag{5}$$

with $\mathscr{Z} = \{\boldsymbol{v} \in \mathbb{R}^N | \mathbf{S}\boldsymbol{v} = 0\}$; in particular, since $\mathscr{D} \cap \mathscr{Z} = \emptyset$, the gradient of R at $\boldsymbol{v} \in \mathscr{D}$ is given by $\nabla R(\boldsymbol{v}) = 2c_{\boldsymbol{v}}\mathbf{S}(c_{\boldsymbol{v}}\boldsymbol{v} - \mathbb{1}) = 2c_{\boldsymbol{v}}(c_{\boldsymbol{v}}\mathbf{S}\boldsymbol{v} - \mathfrak{g})$.

Theorem 2.2. *The map R is quasiconvex on $\mathbb{R}^N$, and pseudoconvex on the convex cone $\mathscr{D}$.*

Sketch of proof. For $\boldsymbol{\xi} = (1-\rho)\boldsymbol{v} + \rho\boldsymbol{\eta}$, $\boldsymbol{v}$ and $\boldsymbol{\eta} \in \mathbb{R}^N$, $\rho \in [0, 1]$, there exist $\rho' \in [0, 1]$ and $c \geqslant 0$ such that $c\boldsymbol{\xi} = (1-\rho')c_{\boldsymbol{v}}\boldsymbol{v} + \rho' c_{\boldsymbol{\eta}}\boldsymbol{\eta}$; we then obtain $R(\boldsymbol{\xi}) \leqslant \max\{R(\boldsymbol{v}), R(\boldsymbol{\eta})\}$ from the definition of R and the convexity of D (quasiconvexity).

For $\boldsymbol{v}$ and $\boldsymbol{\eta} \in \mathscr{D}$ such that $\Theta(\boldsymbol{v}; \boldsymbol{\eta}) \geqslant 0$, we have $(\boldsymbol{v}^*\mathbf{S}\mathbb{1})(\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{v}) \geqslant (\boldsymbol{v}^*\mathbf{S}\boldsymbol{v})(\boldsymbol{\eta}^*\mathbf{S}\mathbb{1}) > 0$. By the Cauchy-Schwarz inequality, we then obtain $0 < (\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{v})^2 \leqslant (\boldsymbol{v}^*\mathbf{S}\boldsymbol{v})(\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta})$, and combining these two sets of inequalities gives $(\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{v})^2/(\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta}) \leqslant (\boldsymbol{v}^*\mathbf{S}\mathbb{1})^2/(\boldsymbol{v}^*\mathbf{S}\boldsymbol{v})$, that is $R(\boldsymbol{v}) \leqslant R(\boldsymbol{\eta})$ (pseudoconvexity). □

By definition of R and following [7], for all $\boldsymbol{v} \in \mathbb{R}^N$, we have

$$C_{\mathrm{sp}}(\boldsymbol{v}) \leqslant C_{\mathrm{F}}(\boldsymbol{v}) \leqslant C_{\mathrm{P}}(\boldsymbol{v}) \leqslant C_{\mathrm{PP}}(\boldsymbol{v}) \leqslant R(\boldsymbol{v}) \leqslant D(\boldsymbol{v}), \tag{6}$$

and $R(\boldsymbol{v}) \leqslant \|\mathbf{K}\|_{\mathrm{F}}^2$. For all $i \in [N]$, we also have $C_{\mathrm{PP}}(\boldsymbol{e}_i) = R(\boldsymbol{e}_i)$, $i \in [N]$, with $\boldsymbol{e}_i$ the i-th vector of the canonical basis of $\mathbb{R}^N$. The appearance of the maps D, and R and C_{F} on the convex cone $\mathbb{R}^N_{\geqslant 0}$ is illustrated in Figure 1. The diagonal entries of $\mathbf{K}$ being strictly positive, we have $\mathbb{R}^N_{\geqslant 0}\backslash\{0\} \subset \mathscr{D}$.

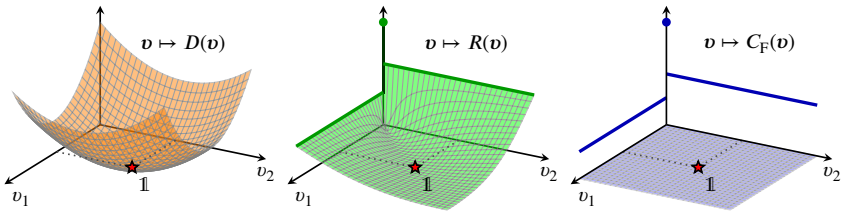

Figure 1: Schematic representation of the maps D, R and C_{F} on $\mathbb{R}^N_{\geqslant 0}$; the red star represents the target selection vector $\mathbb{1} \in \mathbb{R}^N$. The presented graphs correspond to a 2×2 matrix $\mathbf{K}$ such that $\mathbf{K}_{1,1} = 1.225$, $\mathbf{K}_{2,2} = 0.894$ and $\mathbf{K}_{2,1} = 0.316$. In the graphs of R and C_{F}, the point on the vertical axis indicates the value of these maps at $\boldsymbol{v} = 0$ (that is $\|\mathbf{K}\|_{\mathrm{F}}^2$), and the bold lines indicate the constant values taken by the maps along the horizontal axes.

For two selection vectors $\boldsymbol{v}$ and $\boldsymbol{\eta} \in \mathscr{D}$, we set $\mathcal{P}_{\boldsymbol{v}}[\boldsymbol{\eta}] = \boldsymbol{v}(\boldsymbol{v}^*\mathbf{S}\boldsymbol{\eta})/(\boldsymbol{v}^*\mathbf{S}\boldsymbol{v}) \in \mathbb{R}^N$. If $\Theta(\boldsymbol{v}; \boldsymbol{\eta}) < 0$ and $\Theta(\boldsymbol{\eta}; \boldsymbol{v}) \leqslant 0$, then the function $\rho \mapsto R\big(\boldsymbol{v} + \rho(\boldsymbol{\eta} - \boldsymbol{v})\big)$, $\rho \in [0, 1]$, is minimum at $\rho = r \in (0, 1]$, with

$$r = \frac{(\boldsymbol{v}^*\mathbf{S}\boldsymbol{v})\mathfrak{g}^*(\boldsymbol{\eta} - \mathcal{P}_{\boldsymbol{v}}[\boldsymbol{\eta}])}{(\boldsymbol{v}^*\mathbf{S}\boldsymbol{v})\mathfrak{g}^*(\boldsymbol{\eta} - \mathcal{P}_{\boldsymbol{v}}[\boldsymbol{\eta}]) + (\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta})\mathfrak{g}^*(\boldsymbol{v} - \mathcal{P}_{\boldsymbol{\eta}}[\boldsymbol{v}])}, \tag{7}$$

providing the optimal step size for the minimisation of R via a line search from $\boldsymbol{v}$ in the direction $\boldsymbol{\eta} - \boldsymbol{v}$; this descent leads to the improvement

$$\mathcal{I}(\boldsymbol{v}; \boldsymbol{\eta}) = R(\boldsymbol{v}) - R\big(\boldsymbol{v} + r(\boldsymbol{\eta} - \boldsymbol{v})\big) = \big(\mathfrak{g}^*(\boldsymbol{\eta} - \mathcal{P}_{\boldsymbol{v}}[\boldsymbol{\eta}])\big)^2 \big/ \big(\boldsymbol{\eta}^* \mathbf{S}(\boldsymbol{\eta} - \mathcal{P}_{\boldsymbol{v}}[\boldsymbol{\eta}])\big) \geqslant 0. \tag{8}$$

3 Line-search-based sequential sampling

For $\mathbf{f} = (f_i)_{i\in[N]} \in \mathbb{R}^N_{>0}$ and $\varkappa > 0$, we introduce $\mathcal{A}_{\mathbf{f}} = \{\boldsymbol{v} \in \mathbb{R}^N_{\geqslant 0} | \mathbf{f}^* \boldsymbol{v} = \varkappa\} \subset \mathscr{D}$. The set $\mathcal{A}_{\mathbf{f}}$ is convex, and its extreme points are the vectors $\{\boldsymbol{\xi}_i\}_{i\in[N]}$, with $\boldsymbol{\xi}_i = \varkappa \boldsymbol{e}_i / f_i \in \mathbb{R}^N_{\geqslant 0}$. Due to the invariance under rescaling of R, we may without loss of generality set $\varkappa = 1$. Hereafter, we describe a sequential column-sampling procedure based on the minimisation of R over $\mathcal{A}_{\mathbf{f}}$ via line search with optimal step size and Frank-Wolfe (FW) direction (c.f. conditional gradient). Many variants of the proposed algorithm may be considered, see for instance Remarks 3.2 and 3.3. We may remark that the selection vector $\boldsymbol{v}^\star = \varkappa \mathbb{1}/(\mathbf{f}^* \mathbb{1}) \in \mathcal{A}_{\mathbf{f}}$ verifies $R(\boldsymbol{v}^\star) = 0$.

The procedure is initialised at $\boldsymbol{v}^{(1)} = \boldsymbol{\xi}_b \in \mathcal{A}_{\mathbf{f}}$, with

$$b \in \arg\min_{i\in[N]} R(\boldsymbol{\xi}_i) = \arg\max_{i\in[N]} \mathfrak{g}_i^2 / \mathbf{S}_{i,i} \text{ (with } \mathfrak{g}_i = \boldsymbol{e}_i^* \mathfrak{g} \text{ the } i\text{-th entry of } \mathfrak{g} = \mathbf{S}\mathbb{1}), \tag{9}$$

and the selection vector at step $q \in \mathbb{N}$ is denoted by $\boldsymbol{v}^{(q)} \in \mathcal{A}_{\mathbf{f}}$. An iteration of our sampling procedure consists of selecting a descent direction $\boldsymbol{\xi}_u - \boldsymbol{v}^{(q)}$, with $u \in [N]$ such that $\Theta(\boldsymbol{v}^{(q)}; \boldsymbol{\xi}_u) < 0$, and of next performing a line search with the optimal step size r given in (7). Such an iteration results in increasing the weight of the u-th component of $\boldsymbol{v}^{(q)}$ while proportionally decreasing the weights of all its other components. As descent direction, we consider the FW direction $\boldsymbol{\xi}_u - \boldsymbol{v}^{(q)}$, with

$$u \in \arg\min_{i\in[N]} \Theta(\boldsymbol{v}^{(q)}; \boldsymbol{\xi}_i) = \arg\min_{i\in[N]} [\nabla R(\boldsymbol{v}^{(q)})]_i / f_i. \tag{10}$$

The initialisation of the descent via (9) ensures that if $\Theta(\boldsymbol{v}^{(q)}; \boldsymbol{\xi}_i) < 0$, $i \in [N]$, then $\Theta(\boldsymbol{\xi}_i; \boldsymbol{v}^{(q)}) < 0$, so that the descent necessarily occurs in the framework of (7).

A pseudocode of the procedure is given in Algorithm 1. The algorithm produces a sequence $\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \cdots$ of selection vectors with increasing support. At stage $q \in \mathbb{N}$, the number m_q of nonnull entries of $\boldsymbol{v}^{(q)}$ verifies $m_q \leqslant \min(q, N)$, so that early stopping of the procedure ensures sparsity of the resulting selection vector. The algorithm stops if $R(\boldsymbol{v}^{(q)}) = 0$ (that is, if $\boldsymbol{v}^{(q)}$ minimises R over $\mathcal{A}_{\mathbf{f}}$; by pseudoconvexity, this condition is equivalent to the condition $\nabla R(\boldsymbol{v}^{(q)}) = 0$), or when $q = Q$, with $Q \in \mathbb{N}$ a given maximum number of iterations (different stopping rules could be considered).

Algorithm 1: Column sampling via line search with FW direction and optimal step size.

Input: matrix $\mathbf{S}$; vector $\mathbf{f}$; maximum number of iterations $Q \in \mathbb{N}$;

Initialisation: compute $b \in [N]$ using (9); set $q = 1$; $\boldsymbol{v}^{(1)} = \boldsymbol{\xi}_b$ and $I_{\boldsymbol{v}^{(1)}} = \{b\}$;

while $q < Q$ *and* $R(\boldsymbol{v}^{(q)}) > 0$ **do**

  compute $u \in [N]$ using (10) with $\boldsymbol{v} = \boldsymbol{v}^{(q)}$;

  compute the optimal step size r from (7) with $\boldsymbol{\eta} = \boldsymbol{\xi}_u$;

  set $\boldsymbol{v}^{(q+1)} = (1 - r)\boldsymbol{v}^{(q)} + r\boldsymbol{\xi}_u$ and $I_{\boldsymbol{v}^{(q+1)}} = I_{\boldsymbol{v}^{(q)}} \cup \{u\}$; increment: $q \leftarrow q + 1$;

Output: subset $I_{\boldsymbol{v}^{(q)}} \subseteq [N]$;

Remark 3.1. The implementation of Algorithm 1 requires the preliminary computation of the target potential $\mathfrak{g} = \mathbf{S}\mathbb{1}$, with complexity $\mathcal{O}(N^2)$; see Remark 2.4. Once $\mathfrak{g}$ is known, each iteration of Algorithm 1 has complexity $\mathcal{O}(N)$. For $q \in \mathbb{N}$, we have

$$\mathbf{S}\boldsymbol{v}^{(q+1)} = (1 - r)\mathbf{S}\boldsymbol{v}^{(q)} + r(\varkappa / f_u)\mathbf{S}_{\bullet,u},$$

so that sparse updates of the terms $\mathbf{S}\boldsymbol{v}$, $\boldsymbol{v}^*\mathbf{S}\boldsymbol{v}$ and $\mathfrak{g}^*\boldsymbol{v}$ can be easily implemented; furthermore, each iteration of Algorithm 1 only requires access to a single column of $\mathbf{S}$. ◁

In view of (10), the sequences of subsets $I_{\boldsymbol{v}^{(1)}} \subseteq I_{\boldsymbol{v}^{(2)}} \subseteq \cdots \subseteq [N]$ generated by Algorithm 1 depend on the choice of the *restriction vector* $\mathbf{f}$. Our experiments suggest that considering $\mathbf{f} = \mathrm{diag}(\mathbf{K})$, the diagonal of $\mathbf{K}$, appears as a relevant choice (in contrast and for instance, it seems that considering $\mathbf{f} = \mathfrak{g}$ should be avoided). A variant of Algorithm 1 returning sequences of subsets that are independent of the choice of $\mathbf{f}$ is described in Remark 3.2.

Remark 3.2. Instead of considering the steepest conditional descent (10), we may combine the information provided by (5) and (8) to characterise the conditional descent directions inducing the best one-step-ahead improvements. In Algorithm 1, we may hence replace the FW direction (10) by the *best-improvement* (BI) direction

$$u \in \arg\max_{i \in G_{\boldsymbol{\upsilon}}} \mathcal{I}(\boldsymbol{\upsilon}; \boldsymbol{\xi}_i), \text{ with } G_{\boldsymbol{\upsilon}} = \{i \in [N] | [\nabla R(\boldsymbol{\upsilon})]_i < 0\}.$$

The complexity of each iteration of the BI variant of Algorithm 1 is still $\mathcal{O}(N)$; however, in comparison to FW, the resulting procedure is costlier as it requires, in addition to the gradient of R, the computation of the relevant improvement scores. The sequences of subsets produced by this algorithm are independent of the choice of the restriction vector $\mathbf{f}$. ◁

Remark 3.3. For a subset $I \subseteq [N]$ of size m, we denote by $\check{\boldsymbol{\upsilon}}(I) \in \mathbb{R}^N_{\geqslant 0}$ the selection vector minimising D over the set of all nonnegative selection vectors $\boldsymbol{\upsilon} \in \mathbb{R}^N_{\geqslant 0}$ such that $I_{\boldsymbol{\upsilon}} \subseteq I$; the non-trivial entries $[\check{\boldsymbol{\upsilon}}(I)]_I$ of $\check{\boldsymbol{\upsilon}}(I)$ are given by a solution to the quadratic program (QP) associated with the minimisation of the function $\boldsymbol{x} \mapsto \boldsymbol{x}^* \mathbf{S}_{I,I} \boldsymbol{x} - 2\mathfrak{g}_I^* \boldsymbol{x}$ over $\mathbb{R}^m_{\geqslant 0}$. The rescaled selection vector $\boldsymbol{\upsilon}(I) = \varkappa \check{\boldsymbol{\upsilon}}(I)/(\mathbf{f}^* \check{\boldsymbol{\upsilon}}(I)) \in \mathcal{A}_{\mathbf{f}}$ then minimises R over the set of all selection vectors $\boldsymbol{\upsilon} \in \mathcal{A}_{\mathbf{f}}$ such that $I_{\boldsymbol{\upsilon}} \subseteq I$. In Algorithm 1 and its BI variant, at all iterations $q \in \mathbb{N}$, instead of performing a descent with optimal step size, we may then set $\boldsymbol{\upsilon}^{(q+1)} = \boldsymbol{\upsilon}(I_{\boldsymbol{\upsilon}^{(q)}} \cup \{u\})$; we refer to this modified update rule as *weight optimisation* (WO). In terms of numerical complexity and in comparison to a descent with optimal step size, at step q, the WO variants involve the computation of a solution to a QP over $\mathbb{R}^{m_q+1}$ (and we may use $\boldsymbol{\upsilon}^{(q)}$ as a warm start). ◁

4 Numerical experiments

We now illustrate the behaviour of Algorithm 1 and of its variants described in Remarks 3.2 and 3.3. To assess the efficiency of the Nyström approximation induced by a subset $I \subseteq [N]$ of size $m \leqslant N$, we introduce the *approximation factors* (see e.g. [3])

$$\mathscr{E}_{\mathrm{P}}(I) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{HS}(\mathcal{H})}}{\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{HS}(\mathcal{H})}}, \mathscr{E}_{\mathrm{PP}}(I) = \frac{\|\mathbf{K} - P_I \mathbf{K} P_I\|_{\mathrm{HS}(\mathcal{H})}}{\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{HS}(\mathcal{H})}} \text{ and } \mathscr{E}_{\mathrm{X}}(I) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{X}}}{\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{X}}}, \tag{11}$$

$\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$, where $\mathbf{K}^\star_m$ is an optimal rank-m approximation of $\mathbf{K}$ (that is, an approximation obtained by spectral truncation). The values of the approximation factors are necessarily larger than or equal to 1, and the smaller the value, the more accurate the approximation.

Remark 4.1. Denoting by $\lambda_1 \geqslant \cdots \geqslant \lambda_N \geqslant 0$ the eigenvalues of $\mathbf{K}$ (repeated with multiplicity), for all $m < N$, we have $\|\mathbf{K} - \mathbf{K}^\star_m\|^2_{\mathrm{HS}(\mathcal{H})} = \|\mathbf{K} - \mathbf{K}^\star_m\|^2_{\mathrm{F}} = \sum_{l=m+1}^N \lambda_l^2$, $\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{tr}} = \sum_{l=m+1}^N \lambda_l$ and $\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{sp}} = \lambda_{m+1}$. ◁

We implement Algorithm 1 (referred to as FW, for short) and its BI variant (referred to as BI); in addition to the optimal-step-size update rule, for both the FW and BI descent directions, we also implement the WO update rule (the resulting procedures are referred to as FW-WO and BI-WO). The affine restrictions are defined with $\mathbf{f} = \mathrm{diag}(\mathbf{K})$ and $\varkappa = 1$. Due to the specificity of our sampling procedures (which are based on early stopping of line-search-based strategies with sparse initialisations and sparse descent directions), in all our experiments, we placed a special emphasis on approximations involving a relatively small number of columns; in this range, we in particular have $m_q = q$ (with m_q the size of the support of $\boldsymbol{\upsilon}^{(q)}$). We compare our procedures with random sampling with respect to uniform weights and weights proportional to the diagonal of $\mathbf{K}$, *leverage-score-based* random sampling, and *determinantal-point-process-based* (DPP-based) random sampling; see for instance [5, 11, 9, 13, 17, 4] for an overview.

4.1 Random PSD matrix

We consider a random PSD matrix $\mathbf{K} \in \mathbb{C}^{N \times N}$, with $N = 1{,}500$; the eigenvalues of $\mathbf{K}$ are independent realisations of a log-normal distribution ($\mu = -2.5$ and $\sigma = 3$), and a set of associated eigenvectors is defined using a random unitary matrix (multiplication-invariant Haar measure; see [14]).

The evolution of the error maps R and C_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}, \mathrm{PP}\}$, during the 100 first iterations of Algorithm 1 and its BI variant is illustrated in Figure 2. Following (6), the evolution of of C_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}, \mathrm{PP}\}$, is

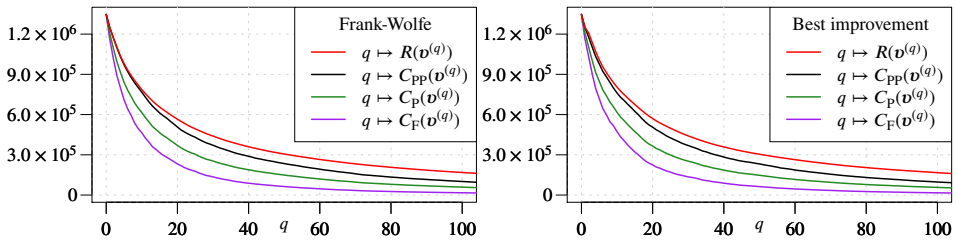

Figure 2: For a (complex) random PSD matrix with $N = 1{,}500$, evolution of the error maps R and C_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}, \mathrm{PP}\}$, during the 100 first iterations of Algorithm 1 (left) and its BI variant (right).

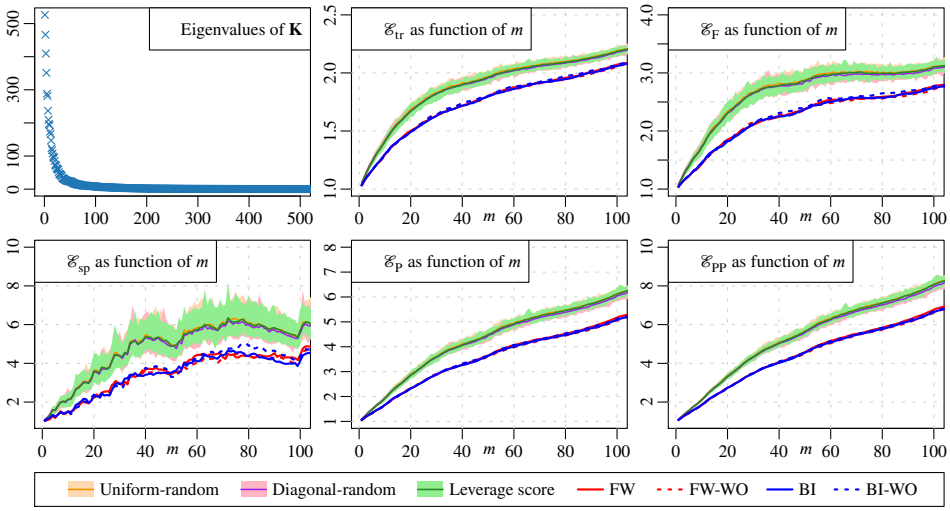

Figure 3: For a (complex) random PSD matrix with $N = 1{,}500$, and for various sequential sampling strategies, evolution of the five approximation factors (11) as functions of the number of columns m. The 500 largest eigenvalues of $\mathbf{K}$ are also represented. For the stochastic strategies, we present the median, minimum and maximum of the approximation factors over 100 repetitions.

bounded by the decay of R (these four error maps are considered since they take the same value at $\upsilon = 0$). We observe a strong similarity between the evolution of these maps, further supporting the use of R as surrogate error map for Nyström approximation.

We then implement various sampling strategies, and compare the evolution of the five approximation factors $\mathscr{E}_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}, \mathrm{P}, \mathrm{PP}\}$, as functions of m (number of columns). For the stochastic strategies, 100 repetitions are performed. The result are presented in Figure 3. In the considered regime (that is, $m \ll m$), we observe that independently of the approximation factor considered, the Nyström approximations induced by Algorithm 1 and its variants are more accurate than the ones obtained using uniform random sampling, diagonal random sampling or leverage-score-based random sampling.

4.2 Abalone data set

We consider the Abalone data set (UCI Machine Learning Repository; see [6]). Two entries of the data set appearing as outliers are removed, and the features are standardised; the resulting data set consists of $N = 4{,}175$ points in $\mathbb{R}^d$, with $d = 8$. We use this data set and a squared-exponential kernel $K(x, x') = e^{-\gamma \|x - x'\|^2}$, $x, x' \in \mathbb{R}^d$ and $\gamma > 0$ (with $\|.\|$ the Euclidean norm of $\mathbb{R}^d$), to generate a PSD matrix $\mathbf{K}$. To illustrate the impact of the decay of the spectrum of $\mathbf{K}$ on the sampling process, we consider different values of γ, namely $\gamma = 0.1, 0.25$ and 1, chosen so that the the eigenvalues of $\mathbf{K}$ exhibit relatively steep, moderate and shallow decays, respectively; see Figure 4.

The accuracy of the approximations induced by the four variants of Algorithm 1 (namely FW, BI, FW-WO, BI-WO) is compared with the accuracy of the approximations obtained via uniform random sampling, leverage-score-based random sampling and k-DPP-based random sampling. The

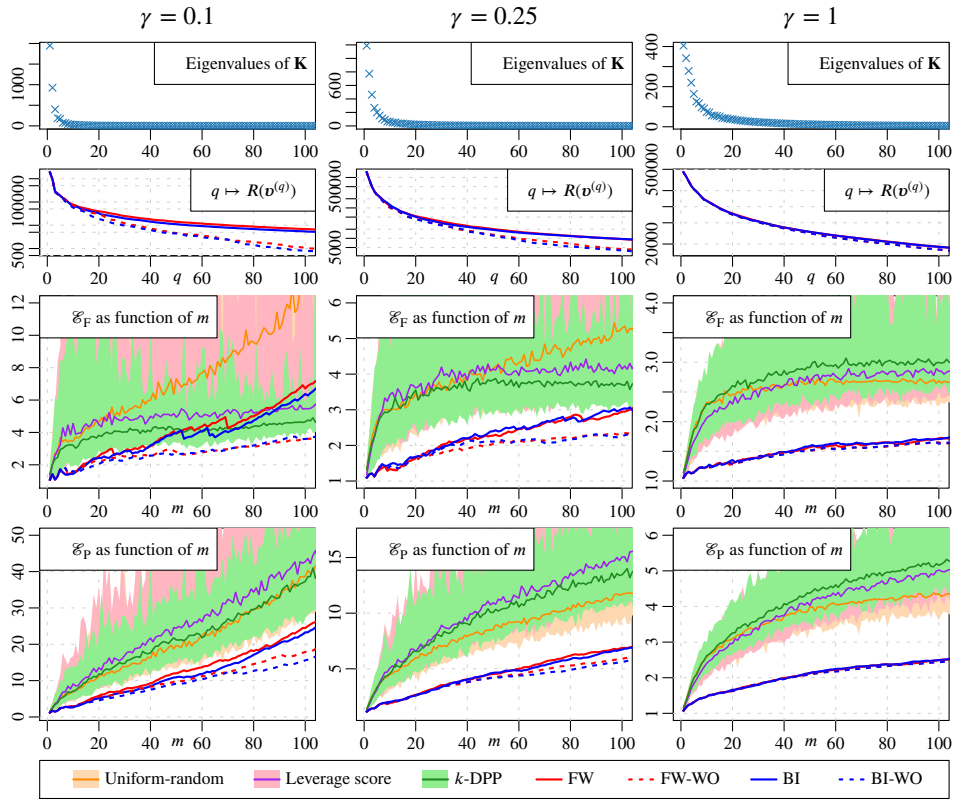

Figure 4: For kernel matrices defined from the Abalone data set and squared exponential kernels, evolution of the approximation factors $\mathscr{E}_{\mathrm{F}}$ and $\mathscr{E}_{\mathrm{P}}$ as functions of the number of columns m. Each column in the figure corresponds to a different value of the kernel parameter γ. For each γ, the 100 largest eigenvalues of $\mathbf{K}$ are displayed, together with the decay, in logarithmic scale, of the error map R during the 100 first iterations of the FW and BI variants of Algorithm 1, with both optimal-step-size and WO update rules. The evolution of $\mathscr{E}_{\mathrm{F}}$ and $\mathscr{E}_{\mathrm{P}}$ are represented for the four variants of Algorithm 1, as well as for random sampling strategies based on uniform weights, leverage scores and k-DPPs. For stochastic methods, the solid line represents the median over 100 repetitions, and the shaded regions indicate the corresponding maximum and minimum values.

experiments involving random sampling are repeated 100 times. The result are presented in Figure 4, where we display the evolution of the approximation factors $\mathscr{E}_{\mathrm{F}}$ and $\mathscr{E}_{\mathrm{P}}$ up to $m = 100$ (the evolution of the other approximation factors in provided in appendix; in terms of behaviour, $\mathscr{E}_{\mathrm{tr}}$ and $\mathscr{E}_{\mathrm{sp}}$ appear closely related to $\mathscr{E}_{\mathrm{F}}$, while $\mathscr{E}_{\mathrm{PP}}$ shows similarities with $\mathscr{E}_{\mathrm{P}}$). In comparison to the considered random-sampling procedures, we observe that Algorithm 1 and its variants lead to accurate approximations, especially in the range corresponding to the significant eigenvalues of $\mathbf{K}$.

After a certain number of iterations (which appears to be related to the decay of the spectrum of $\mathbf{K}$), the accuracy of the approximations induced by Algorithm 1 and its BI variant deteriorate (this is especially visible for $\gamma = 0.1$). This behaviour is related to the relatively slow asymptotic convergence of conditional gradient descents (see e.g. [2, 12]); the deterioration is stronger for $\mathscr{E}_{\mathrm{F}}$ than for $\mathscr{E}_{\mathrm{P}}$, and the WO update rule appears to be able to mitigate this drop-off in accuracy.

4.3 HIGGS data set

We now illustrate the ability of the proposed approach to tackle matrices of very large scale. We consider the HIGGS dataset (UCI Machine Learning Repository; see [6]), consisting of $N = 11{,}000{,}000$ points in $\mathbb{R}^d$, with $d = 21$; all the features are standardised. To define a PSD matrix $\mathbf{K}$, we use a squared-exponential kernel (same expression as in Section 4.2) with $\gamma = 0.2$. In double-precision floating-point format, storing all the entries of $\mathbf{K}$ or $\mathbf{S}$ would require more than 968 terabytes of

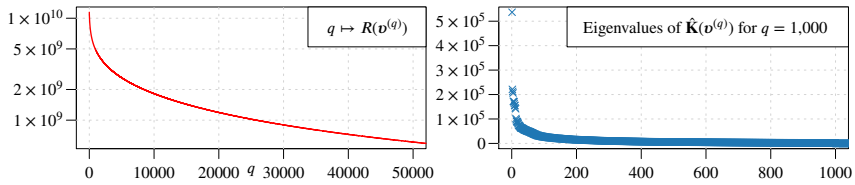

Figure 5: For the HIGGS data set, decay of the the error map R during the 50,000 first iterations of Algorithm 1 (logarithmic scale). The main eigenvalues of the Nyström approximation of $\mathbf{K}$ obtained at $q = 1{,}000$ are also presented.

memory; as an alternative, rather than being stored, the entries of the matrix $\mathbf{S}$ are computed on demand from the data set and the kernel (*on-the-fly evaluation*).

In Figure 5, we display the decay of the error map R during the first 50,000 iterations of Algorithm 1. The inequalities (6) ensures that the evolution of the error maps C_{X}, $\mathrm{X} \in \{\mathrm{sp}, \mathrm{F}, \mathrm{P}, \mathrm{PP}\}$ is bounded by the decay of R (see Figure 2 for an illustration) We also present the eigenvalues of the Nyström approximation $\hat{\mathbf{K}}(\boldsymbol{\upsilon}^{(q)})$ of $\mathbf{K}$ for $q = 1{,}000$; this approximation involves $m_q = 1{,}000$ columns of $\mathbf{K}$.

Remark 4.2. For the HIGGS data set, both Algorithm 1 and a procedure to obtain the target potential $\mathfrak{g} = \mathbf{S}\mathbb{1}$ were implemented in `C`; both relied on on-the-fly evaluations of the entries of $\mathbf{S}$. Using 50 CPU threads (AMD Ryzen Threadripper 3990X @ 2.9GHz), the computation of the target potential $\mathfrak{g}$ took approximately 63.5 hours. On a single thread, each iteration of Algorithm 1 then took approximately 0.234 seconds (average over the 500,000 first iterations of the algorithm). ◁

5 Conclusion

We presented a class of gradient-based sequential sampling strategies for Nyström approximation which leverages the properties of the differentiable surrogate error map R. The proposed strategies are based on early stopping of line-search-type procedures with sparse initialisations and sparse descent directions; as such, they are primarily intended to be used to extract relatively small samples of columns (that is, $m \ll N$). In view of our experiments, and especially for the optimal-step-size update rule, the range in which these strategies are able to maintain a high level of accuracy appears to be related to the decay of the eigenvalues of $\mathbf{K}$; gaining a deeper understanding of the mechanisms at play could help to further improve the operating framework of this type of strategies.

From a numerical standpoint, the main bottleneck of the proposed framework is the computation of the target potential $\mathfrak{g} \in \mathbb{R}^N$ (quadratic complexity); this operation can nevertheless be easily parallelised. Once $\mathfrak{g}$ is known, the complexity of each iteration of the considered strategies is linear in N (and the space complexity is also linear). Further, as they involves basic algebraic operations, to improve their efficiency, Algorithm 1 and its variants could be implemented using GPUs; stochastic approximations of $\mathfrak{g}$ may also be considered. In complement to sequential sampling, other types of strategies leveraging the properties of the energy setting may be considered, such as regularisation-based approaches and, for kernel matrices specifically, particle-flow-based approaches; see e.g. [8, 10].

References

[1] Ahmed Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. *Advances in Neural Information Processing Systems*, 28:775–783, 2015.

[2] Michael D. Canon and Clifton D. Cullum. A tight upper bound on the rate of convergence of the Frank-Wolfe algorithm. *SIAM Journal on Control*, 6(4):509–516, 1968.

[3] Michal Derezinski, Rajiv Khanna, and Michael W. Mahoney. Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nyström method. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[4] Michal Derezinski and Michael W. Mahoney. Determinantal point processes in randomized numerical linear algebra. *Notices of the American Mathematical Society*, 68(1):34–45, 2021.

[5] Petros Drineas and Michael W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

[6] Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2019.

[7] Bertrand Gauthier. Isometric representation of integral operators with positive-semidefinite kernels. `https://hal.science/hal-03848105v2/document`, 2023.

[8] Bertrand Gauthier and Johan Suykens. Optimal quadrature-sparsification for integral operator approximation. *SIAM Journal on Scientific Computing*, 40:A3636–A3674, 2018.

[9] Alex Gittens and Michael W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *Journal of Machine Learning Research*, 17:1–65, 2016.

[10] Matthew Hutchings and Bertrand Gauthier. Local optimisation of Nyström samples through stochastic gradient descent. In *Machine Learning, Optimization, and Data Science - LOD 2022*, 2023.

[11] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 13:981–1006, 2012.

[12] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. *Advances in Neural Information Processing Systems*, 28:496–504, 2015.

[13] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast DPP sampling for Nyström with application to kernel methods. In *International Conference on Machine Learning*, volume 48, pages 2061–2070. PMLR, 2016.

[14] Francesco Mezzadri. How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54(5):592–604, 2007.

[15] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: a review and beyond. *Foundations and Trends® in Machine Learning*, 10:1–141, 2017.

[16] Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

[17] Edouard Pauwels, Francis Bach, and Jean-Philippe Vert. Relating leverage scores and density using regularized Christoffel functions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[18] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT press, Cambridge, MA, 2006.

[19] Irene Rodriguez-Lujan, Charles Elkan, Carlos Santa Cruz Fernández, Ramón Huerta, et al. Quadratic programming feature selection. *Journal of Machine Learning Research*, 11:1491–1516, 2010.

[20] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R.G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[21] Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35:363–417, 2012.

[22] Nicholas Sterge, Bharath Sriperumbudur, Lorenzo Rosasco, and Alessandro Rudi. Gain with no pain: efficiency of kernel-PCA by Nyström sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 3642–3652. PMLR, 2020.

[23] Shusen Wang, Alex Gittens, and Michael W. Mahoney. Scalable kernel K-means clustering with Nyström approximation: relative-error bounds. *The Journal of Machine Learning Research*, 20(1):431–479, 2019.

[24] Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems*, 13:682–688, 2000.

A Details of the proofs

Complement to the proof of Theorem 2.1. As illustrated in the main body of the paper, the result follows directly from the fact that if $J \subseteq I \subseteq [N]$, then $\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{X}} \leqslant \|\mathbf{K} - \hat{\mathbf{K}}(J)\|_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$; we shall thus merely focus on proving these inequalities. We denote by $\mathcal{H}_{0I}$ the orthogonal complement of $\mathcal{H}_I$ in $\mathcal{H}$, so that $\mathbf{K} - \hat{\mathbf{K}}(I) = P_{0I}\mathbf{K}$, with P_{0I} the orthogonal projection from $\mathcal{H}$ onto $\mathcal{H}_{0I}$; we similarly introduce the subspace $\mathcal{H}_{0J}$ and the orthogonal projection P_{0J}. We have $\mathcal{H}_{0I} \subseteq \mathcal{H}_{0J}$, and we also denote by $\mathcal{H}_e$ the orthogonal complement of $\mathcal{H}_{0I}$ in $\mathcal{H}_{0J}$ (so that $P_{0J} = P_{0I} + P_e$).

We first consider the trace norm. Let $\mathrm{HS}(\mathcal{E}, \mathcal{H})$ be the Hilbert space of all HS operators from $\mathcal{E}$ to $\mathcal{H}$. Following Remark 2.3 and noticing that $\langle P_e\mathbf{K} \mid P_{0I}\mathbf{K}\rangle_{\mathrm{HS}(\mathcal{E},\mathcal{H})} = 0$, we have

$$\|\mathbf{K} - \hat{\mathbf{K}}(J)\|_{\mathrm{tr}} = \|P_{0J}\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{E},\mathcal{H})} = \|P_{0I}\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{E},\mathcal{H})} + \|P_e\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{E},\mathcal{H})} \geqslant \|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{tr}},$$

as expected.

We now consider the Frobenius and spectral norms, and we denote by $\mathrm{HS}(\mathcal{E})$ the Hilbert space of all HS operators on $\mathcal{E}$. We recall that if P is an orthogonal projection on $\mathcal{H}$, then the PSD operator on $\mathcal{E}$ related to $P\mathbf{K}$ and the PSD operator $\mathcal{H}$ related to $P\mathbf{K}P$ have the same nonnegative eigenvalues (this for instance follows by noticing that the natural embedding of $\mathcal{H}$ in $\mathcal{E}$ is HS, and that the considered operators are both composites of this embedding; see e.g. [21, 7]); we then in particular have $\|P\mathbf{K}\|_{\mathrm{HS}(\mathcal{E})} = \|P\mathbf{K}P\|_{\mathrm{HS}(\mathcal{H})}$. As $\mathcal{H}_{0I}$ and $\mathcal{H}_e$ are orthogonal in $\mathcal{H}$, the operators related to $P_{0I}\mathbf{K}P_{0I}$, $P_e\mathbf{K}P_e$, $P_{0I}\mathbf{K}P_e$ and $P_e\mathbf{K}P_{0I}$ are orthogonal in $\mathrm{HS}(\mathcal{H})$; for the Frobenius norm, we obtain

$$\begin{aligned}
\|\mathbf{K} - \hat{\mathbf{K}}(J)\|^2_{\mathrm{F}} &= \|P_{0J}\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{E})} = \|P_{0J}\mathbf{K}P_{0J}\|^2_{\mathrm{HS}(\mathcal{H})} \\
&= \|P_{0I}\mathbf{K}P_{0I}\|^2_{\mathrm{HS}(\mathcal{H})} + \|P_e\mathbf{K}P_e\|^2_{\mathrm{HS}(\mathcal{H})} + \|P_{0I}\mathbf{K}P_e\|^2_{\mathrm{HS}(\mathcal{H})} + \|P_e\mathbf{K}P_{0I}\|^2_{\mathrm{HS}(\mathcal{H})} \\
&\geqslant \|P_{0I}\mathbf{K}P_{0I}\|^2_{\mathrm{HS}(\mathcal{H})} = \|P_{0I}\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{E})} = \|\mathbf{K} - \hat{\mathbf{K}}(I)\|^2_{\mathrm{F}},
\end{aligned}$$

providing the expected inequality. For the spectral norm, as $\mathcal{H}_{0I} \subseteq \mathcal{H}_{0J}$, we also get

$$\begin{aligned}
\|\mathbf{K} - \hat{\mathbf{K}}(J)\|_{\mathrm{sp}} &= \max\{\langle \mathbf{v} \mid P_{0J}\mathbf{K}\mathbf{v}\rangle_{\mathcal{E}} | \mathbf{v} \in \mathcal{E}, \|\mathbf{v}\|_{\mathcal{E}} = 1\} \\
&= \max\{\langle \boldsymbol{h} \mid P_{0J}\mathbf{K}P_{0J}\boldsymbol{h}\rangle_{\mathcal{H}} | \boldsymbol{h} \in \mathcal{H}, \|\boldsymbol{h}\|_{\mathcal{H}} = 1\} \\
&= \max\{\langle P_{0J}\boldsymbol{h} \mid \mathbf{K}P_{0J}\boldsymbol{h}\rangle_{\mathcal{H}} | \boldsymbol{h} \in \mathcal{H}, \|\boldsymbol{h}\|_{\mathcal{H}} = 1\} \\
&= \max\{\langle \boldsymbol{h} \mid \mathbf{K}\boldsymbol{h}\rangle_{\mathcal{H}} | \boldsymbol{h} \in \mathcal{H}_{0J}, \|\boldsymbol{h}\|_{\mathcal{H}} = 1\} \\
&\geqslant \max\{\langle \boldsymbol{h} \mid \mathbf{K}\boldsymbol{h}\rangle_{\mathcal{H}} | \boldsymbol{h} \in \mathcal{H}_{0I}, \|\boldsymbol{h}\|_{\mathcal{H}} = 1\} = \|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{sp}},
\end{aligned}$$

as required. □

Proof of Corollary 2.1. We simply need to show that if $J \subseteq I \subseteq [N]$, then

$$\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{HS}(\mathcal{H})} \leqslant \|\mathbf{K} - \hat{\mathbf{K}}(J)\|_{\mathrm{HS}(\mathcal{H})} \quad \text{and} \quad \|\mathbf{K} - P_I\mathbf{K}P_I\|_{\mathrm{HS}(\mathcal{H})} \leqslant \|\mathbf{K} - P_J\mathbf{K}P_J\|_{\mathrm{HS}(\mathcal{H})}.$$

Using the same notations as in the proof of Theorem 2.1 and noticing that $\mathcal{H}_{0I}$ and $\mathcal{H}_e$ are orthogonal in $\mathcal{H}$, we have

$$\|\mathbf{K} - \hat{\mathbf{K}}(J)\|^2_{\mathrm{HS}(\mathcal{H})} = \|P_{0J}\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{H})} = \|P_{0I}\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{H})} + \|P_e\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{H})} \geqslant \|\mathbf{K} - \hat{\mathbf{K}}(I)\|^2_{\mathrm{HS}(\mathcal{H})},$$

as required. Next, if P is an orthogonal projection on $\mathcal{H}$, then (see e.g. [7])

$$\|\mathbf{K} - P\mathbf{K}P\|^2_{\mathrm{HS}(\mathcal{H})} = \|\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{H})} - \|P\mathbf{K}P\|^2_{\mathrm{HS}(\mathcal{H})}. \tag{12}$$

Denoting by $\mathcal{H}_r$ the orthogonal complement of $\mathcal{H}_J$ in $\mathcal{H}_I$ and noticing that the operators related to $P_J\mathbf{K}P_J$, $P_r\mathbf{K}P_r$, $P_J\mathbf{K}P_r$ and $P_r\mathbf{K}P_J$ are orthogonal in $\mathrm{HS}(\mathcal{H})$, we obtain

$$\begin{aligned}
\|P_I\mathbf{K}P_I\|^2_{\mathrm{HS}(\mathcal{H})} &= \|P_J\mathbf{K}P_J\|^2_{\mathrm{HS}(\mathcal{H})} + \|P_r\mathbf{K}P_J\|^2_{\mathrm{HS}(\mathcal{H})} + \|P_J\mathbf{K}P_r\|^2_{\mathrm{HS}(\mathcal{H})} + \|P_r\mathbf{K}P_r\|^2_{\mathrm{HS}(\mathcal{H})} \\
&\geqslant \|P_J\mathbf{K}P_J\|^2_{\mathrm{HS}(\mathcal{H})},
\end{aligned}$$

giving, in combination with (12), the expected inequality. In both cases, we conclude by following the same reasoning as in the sketch of the proof of Theorem 2.1. □

Complement to the proof of Theorem 2.2. We first discuss the quasiconvexity of R on $\mathbb{R}^N$. As mentioned in the sketch of the proof, for $\boldsymbol{\xi} = \boldsymbol{\upsilon} + \rho(\boldsymbol{\eta} - \boldsymbol{\upsilon})$, $\boldsymbol{\upsilon}$ and $\boldsymbol{\eta} \in \mathbb{R}^N$, $\rho \in [0, 1]$, there exists $c \geqslant 0$ and $\rho' \in [0, 1]$ such that $c\boldsymbol{\xi} = (1 - \rho')c_{\boldsymbol{\upsilon}}\boldsymbol{\upsilon} + \rho' c_{\boldsymbol{\eta}}\boldsymbol{\eta}$; indeed:

- for $\boldsymbol{\upsilon} \notin \mathscr{D}$ and $\boldsymbol{\eta} \notin \mathscr{D}$, the condition is verified for $c = 0$ and for any $\rho' \in [0, 1]$;
- for $\boldsymbol{\upsilon} \notin \mathscr{D}$ and $\boldsymbol{\eta} \in \mathscr{D}$, the condition is verified for $c = 0$ and $\rho' = 0$;
- for $\boldsymbol{\upsilon} \in \mathscr{D}$ and $\boldsymbol{\eta} \notin \mathscr{D}$, the condition is verified for $c = 0$ and $\rho' = 1$;
- for $\boldsymbol{\upsilon} \in \mathscr{D}$ and $\boldsymbol{\eta} \in \mathscr{D}$, we have $\mathrm{coni}\{\boldsymbol{\upsilon}, \boldsymbol{\eta}\} = \mathrm{coni}\{c_{\boldsymbol{\upsilon}}\boldsymbol{\upsilon}, c_{\boldsymbol{\eta}}\boldsymbol{\eta}\}$ (with $\mathrm{coni}\{\boldsymbol{\upsilon}, \boldsymbol{\eta}\}$ the conical hull of $\{\boldsymbol{\upsilon}, \boldsymbol{\eta}\}$), so that $\boldsymbol{\xi} \in \mathrm{coni}\{c_{\boldsymbol{\upsilon}}\boldsymbol{\upsilon}, c_{\boldsymbol{\eta}}\boldsymbol{\eta}\}$ (in this case, $\boldsymbol{\xi} \in \mathscr{D}$, and so $c > 0$).

From the definition of R and the convexity of D, we obtain

$$R(\boldsymbol{\xi}) \leqslant D(c\boldsymbol{\xi}) \leqslant (1 - \rho')D(c_{\boldsymbol{\upsilon}}\boldsymbol{\upsilon}) + \rho' D(c_{\boldsymbol{\eta}}\boldsymbol{\eta}) = (1 - \rho')R(\boldsymbol{\upsilon}) + \rho' R(\boldsymbol{\eta}) \leqslant \max\{R(\boldsymbol{\upsilon}), R(\boldsymbol{\eta})\},$$

so that R is quasiconvex on $\mathbb{R}^N$.

We now prove the pseudoconvexity of R on $\mathscr{D}$. For $\boldsymbol{\upsilon} \in \mathscr{D}$, we have $\boldsymbol{\upsilon}^*\mathbf{S}(c_{\boldsymbol{\upsilon}}\boldsymbol{\upsilon} - \mathbb{1}) = 0$, so that the condition $\Theta(\boldsymbol{\upsilon}; \boldsymbol{\eta}) \geqslant 0$, $\boldsymbol{\eta} \in \mathscr{D}$, reads $\boldsymbol{\eta}^*\mathbf{S}(c_{\boldsymbol{\upsilon}}\boldsymbol{\upsilon} - \mathbb{1}) \geqslant 0$, that is,

$$(\boldsymbol{\upsilon}^*\mathbf{S}\mathbb{1})(\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\upsilon}) \geqslant (\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon})(\boldsymbol{\eta}^*\mathbf{S}\mathbb{1}). \tag{13}$$

As $\boldsymbol{\upsilon}$ and $\boldsymbol{\eta} \in \mathscr{D}$, we have $\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon} > 0$ and $\boldsymbol{\eta}^*\mathbf{S}\mathbb{1} > 0$, and thus, from (13), $\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\upsilon} > 0$. The matrix $\mathbf{S}$ being SPSD, the Cauchy-Schwarz inequality gives $(\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\upsilon})^2 \leqslant (\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon})(\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta})$; combining the CS inequality with (13), we obtain (note that we also have $\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta} > 0$)

$$\frac{(\boldsymbol{\upsilon}^*\mathbf{S}\mathbb{1})^2}{(\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon})^2} \geqslant \frac{(\boldsymbol{\eta}^*\mathbf{S}\mathbb{1})^2}{(\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\upsilon})^2} \geqslant \frac{(\boldsymbol{\eta}^*\mathbf{S}\mathbb{1})^2}{(\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon})(\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta})}.$$

We thus have $(\boldsymbol{\eta}^*\mathbf{S}\mathbb{1})^2/(\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta}) \leqslant (\boldsymbol{\upsilon}^*\mathbf{S}\mathbb{1})^2/(\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon})$, that is $R(\boldsymbol{\upsilon}) \leqslant R(\boldsymbol{\eta})$, and R is therefore pseudoconvex on $\mathscr{D}$. □

B Derivation of the optimal-step-size formula

For $\boldsymbol{\upsilon}$ and $\boldsymbol{\eta} \in \mathscr{D}$ such that $\Theta(\boldsymbol{\upsilon}; \boldsymbol{\eta}) < 0$ and $\Theta(\boldsymbol{\eta}; \boldsymbol{\upsilon}) \leqslant 0$, and from the function $\rho \mapsto R\big(\boldsymbol{\upsilon} + \rho(\boldsymbol{\eta} - \boldsymbol{\upsilon})\big)$, $\rho \in [0, 1]$, we introduce

$$f(x) = -\frac{[a + x(b - a)]^2}{c + x^2(d + c - 2e) + 2x(e - c)} = -\frac{\varphi^2(x)}{\psi(x)}, x \in [0, 1],$$

with $a = \boldsymbol{\upsilon}^*\mathbf{S}\mathbb{1} > 0$, $b = \boldsymbol{\eta}^*\mathbf{S}\mathbb{1} > 0$, $c = \boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon} > 0$, $d = \boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta} > 0$ and $e = \boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\eta}$. We then have

$$f'(x) = \frac{\varphi(x)}{\psi^2(x)}\Big[\big(2x(d + c - 2e) + 2(e - c)\big)\varphi(x) - 2(b - a)\psi(x)\Big], x \in [0, 1].$$

On $\mathbb{R}$, and if $a \neq b$, the function φ vanishes at $x_1 = a/(a - b) \notin [0, 1]$ (as $a > 0$ and $b > 0$). Also, the function

$$x \mapsto \big(2x(d + c - 2e) + 2(e - c)\big)\varphi(x) - 2(b - a)\psi(x)$$

is a polynomial of degree 1 on $\mathbb{R}$ which vanishes at

$$x_2 = \frac{bc - ae}{bc - ae + ad - be} = r,$$

with r given by (7), and the pseudoconvexity of R on $\mathscr{D}$ ensures that $r \in (0, 1]$.

C Abalone data set: additional figure

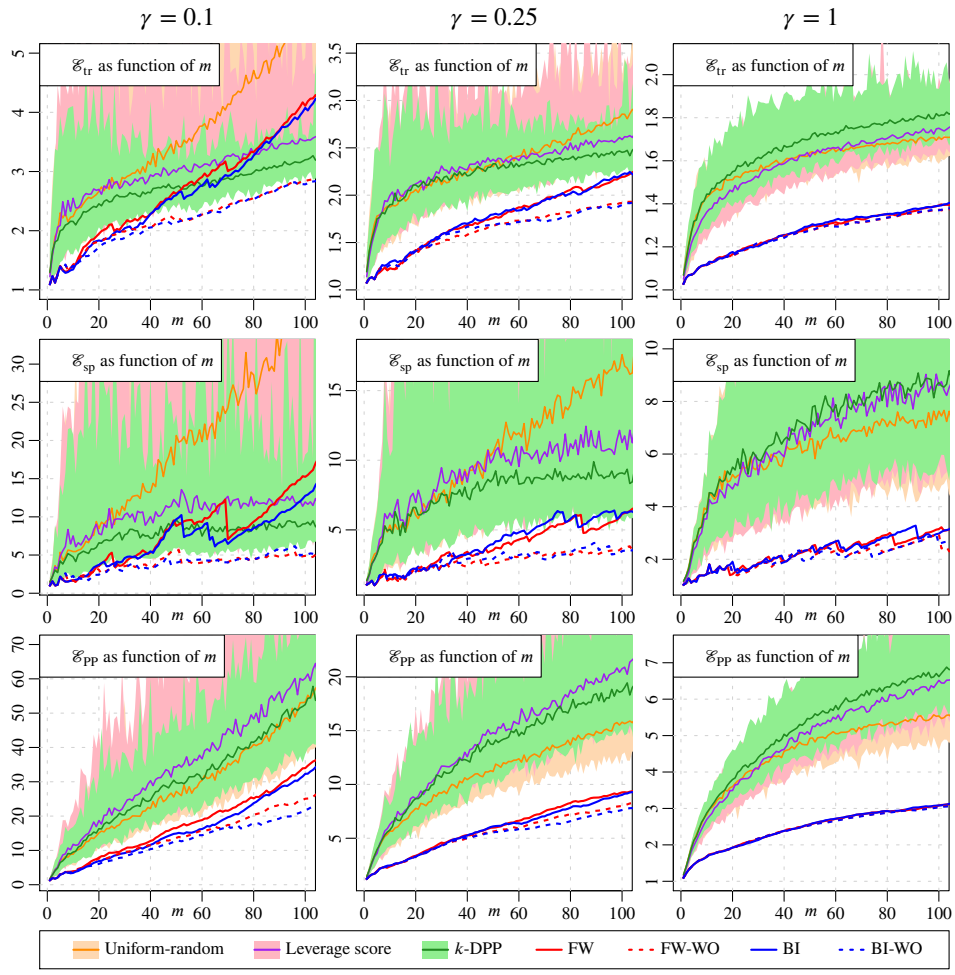

Figure 6: In complement to Figure 4 and for the various sampling strategies considered in Section 4.2, evolution of the approximation factors $\mathscr{E}_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{sp}, \mathrm{PP}\}$, as functions of the number of columns m (Abalone data set and squared-exponential kernel).