CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/171109/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Hutchings, Matthew and Gauthier, Bertrand 2024. Energy-based sequential sampling for low-rank PSD-matrix approximation. SIAM Journal on Mathematics of Data Science

Publishers page:

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

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

Matthew Hutchings* and Bertrand Gauthier†

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 relies on 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. The time complexity of the stochastic variants of the discussed strategies is linear in the order of the considered PSD matrices, and the underlying computations can be easily parallelised.

Key words. Nyström approximation, reproducing kernel Hilbert spaces, differentiable relaxation, generalised convexity, conditional gradient.

MSC codes. 65F55, 46E22, 47B32.

1. Introduction.

The low-rank approximation of matrices through column sampling is a core technique in scientific computing and machine learning. For positive-semidefinite (PSD) matrices, the terminology *Nyström approximation* is often used, and the characterisation of samples of columns leading to accurate approximations is referred to as the *column sampling problem* (CSP); see e.g. [26, 1, 24, 23]. In practical applications, the combinatorial nature of the CSP and the cost inherent to the evaluation of the Nyström 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 [7, 12, 10, 17, 20, 6] and references therein for an overview.

In this work, we describe a class of sequential sampling strategies leveraging the properties of a differentiable pseudoconvex relaxation of the CSP. We characterise samples of columns through the non-zero entries of *selection vectors* (interestingly enough, this alone leads to a convex, but non-differentiable, relaxation of the CSP; see Theorem 2.2); such selection vectors can be regarded as discrete measures, and together with the considered PSD matrix, define integral operators acting on the reproducing kernel Hilbert space (RKHS; see e.g. [19]) defined by the matrix. Following [9, 8], the norm of the corresponding Hilbert-Schmidt (HS) space can be used to discriminate among selection vectors, and enforcing an invariance with respect to the rescaling of selection vectors gives rise to a quasiconvex error map R on the selection-vector space (R is in addition pseudoconvex on a specific convex cone of interest, see Theorem 2.5). The error map R can be minimised through gradient descent, and we describe sequential sampling strategies based on minimisation procedures with sparse initialisation and sparse descent directions (see e.g. [4, 2, 13] for related gradient-based approaches for sampling); sparsity of the samples is enforced by early stopping of the optimisation.

For a $N \times N$ PSD matrix $\mathbf{K}$, the described sampling strategies rely on a vector $\mathfrak{g} \in \mathbb{R}^N$ formed by computing the squared ℓ^2 norm of each row (or column) of $\mathbf{K}$. The time complexity of forming the

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

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

exact *target potential* $\mathfrak{g}$ is therefore quadratic in N; nevertheless, stochastic approximations of $\mathfrak{g}$ can be considered, and the overall time complexity of the proposed strategies is then linear in N (for instance, a strategy with $\mathcal{O}(m^2 + mN + \ell N)$ time complexity is presented, with m the size of the extracted column sample, and where $\ell \ll N$ is a sample-size parameter related to the stochastic approximation of $\mathfrak{g}$). The underlying computations can in addition be easily parallelised.

The manuscript is organised as follows. In Section 2, we describe the overall framework surrounding the considered relaxation of the CSP. In Section 3, we present a class of gradient-based sequential column-sampling strategies, and stochastic variants of these strategies are discussed in Section 4. Section 5 is devoted to numerical experiments, and Section 6 consists of a concluding discussion. Proofs are gathered in appendix, together with some technical results and additional figures.

2. Overall framework and notations. Throughout this note, we use the classical *matrix notation* and identify a vector $\boldsymbol{\alpha} \in \mathbb{C}^N$, $N \in \mathbb{N}$, as the $N \times 1$ column matrix defined by the coefficients of $\boldsymbol{\alpha}$ in the canonical basis $\{\mathbf{e}_i\}_{i\in[N]}$ of $\mathbb{C}^N$; $[N]$ stands for the set of all integers between 1 and N. The conjugate and conjugate-transpose of a matrix $\mathbf{M}$ are denoted by $\overline{\mathbf{M}}$ and $\mathbf{M}^*$, respectively, and $\mathrm{span}\{\mathbf{M}\}$ stands for the linear space spanned by the columns of $\mathbf{M}$. Hermitian forms are assumed to be linear in their second argument.

2.1. Nyström approximation of PSD matrices. Let $\mathbf{K} \in \mathbb{C}^{N\times N}$ be a PSD matrix, with $N \in \mathbb{N}$. For a subset $\mathbb{I} \subseteq [N]$ of size $m \leqslant N$, the *Nyström approximation* of $\mathbf{K}$ induced by $\mathbb{I}$ is the PSD matrix

$$\hat{\mathbf{K}}(\mathbb{I}) = \mathbf{K}_{\bullet,\mathbb{I}}(\mathbf{K}_{\mathbb{I},\mathbb{I}})^{\dagger}\mathbf{K}_{\mathbb{I},\bullet} \in \mathbb{C}^{N\times N}, \tag{2.1}$$

where $\mathbf{K}_{\bullet,\mathbb{I}} \in \mathbb{C}^{N\times m}$ is the matrix defined by the columns of $\mathbf{K}$ with index in $\mathbb{I}$, and where $(\mathbf{K}_{\mathbb{I},\mathbb{I}})^{\dagger}$ is the pseudoinverse of the $m \times m$ principal submatrix of $\mathbf{K}$ defined by $\mathbb{I}$ (and $\mathbf{K}_{\mathbb{I},\bullet} = (\mathbf{K}_{\bullet,\mathbb{I}})^*$ consists of rows of $\mathbf{K}$); see e.g. [7, 21, 12, 10, 5].

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

$$\|\mathbf{K} - \hat{\mathbf{K}}(\mathbb{I})\|_{\mathrm{tr}}, \quad \|\mathbf{K} - \hat{\mathbf{K}}(\mathbb{I})\|_{\mathrm{F}}, \quad \text{or} \quad \|\mathbf{K} - \hat{\mathbf{K}}(\mathbb{I})\|_{\mathrm{sp}}, \tag{2.2}$$

respectively, naturally raising questions related to the characterisation of subsets leading to accurate approximations. In practice, the direct minimisation, as functions of $\mathbb{I}$, of the error norms (2.2) is made difficult by the combinatorial nature of the underlying problems and by the numerical cost inherent to the evaluation of the corresponding norms. The following Remark 2.1 provides an important insight into the theoretical framework surrounding the definition of Nyström approximations and the assessment of their accuracy.

Remark 2.1. The entries of a PSD matrix $\mathbf{K} \in \mathbb{C}^{N\times N}$ characterise the kernel of a RKHS of $\mathbb{C}$-valued functions on $[N]$; see for instance [19, Chapter 2]. This RKHS can be identified with the subspace $\mathcal{H} = \mathrm{span}\{\mathbf{K}\} \subseteq \mathbb{C}^N$ endowed with the inner product

$$\langle \boldsymbol{h} \,|\, \boldsymbol{f} \rangle_{\mathcal{H}} = \boldsymbol{h}^*\mathbf{K}^{\dagger}\boldsymbol{f}, \ \boldsymbol{h} \text{ and } \boldsymbol{f} \in \mathcal{H}.$$

A subset $\mathbb{I} \subseteq [N]$ then defines a closed linear subspace $\mathcal{H}_{\mathbb{I}} = \mathrm{span}\{\mathbf{K}_{\bullet,\mathbb{I}}\}$ of $\mathcal{H}$, and $\hat{\mathbf{K}}(\mathbb{I})$ is the reproducing kernel of $\mathcal{H}_{\mathbb{I}}$. Introducing $P_{\mathbb{I}} = \mathbf{K}_{\bullet,\mathbb{I}}(\mathbf{K}_{\mathbb{I},\mathbb{I}})^{\dagger}\mathbf{I}_{\mathbb{I},\bullet} \in \mathbb{C}^{N\times N}$, with $\mathbf{I}$ the $N \times N$ identity matrix, we indeed

have

$$\hat{\mathbf{K}}(\mathbb{I}) = P_{\mathbb{I}}\mathbf{K} = \mathbf{K}P_{\mathbb{I}}^* = P_{\mathbb{I}}\mathbf{K}P_{\mathbb{I}}^*,$$

and the matrix $P_{\mathbb{I}}$ corresponds to the orthogonal projection from $\mathcal{H}$ onto $\mathcal{H}_{\mathbb{I}}$ (see Remark 2.3), that is

$$\operatorname{span}\{P_{\mathbb{I}}\mathbf{K}\} = \mathcal{H}_{\mathbb{I}}, \quad P_{\mathbb{I}}^2 = P_{\mathbb{I}} \quad \text{and} \quad \langle \boldsymbol{h} \mid P_{\mathbb{I}}\boldsymbol{f}\rangle_{\mathcal{H}} = \langle P_{\mathbb{I}}\boldsymbol{h} \mid \boldsymbol{f}\rangle_{\mathcal{H}}, \, \boldsymbol{h} \text{ and } \boldsymbol{f} \in \mathcal{H}.$$

Denoting by $\mathcal{E}$ the Euclidean Hilbert space $\mathbb{C}^N$ (with inner product $\langle \mathbf{u} \mid \mathbf{v}\rangle_{\mathcal{E}} = \mathbf{u}^*\mathbf{v}$, $\mathbf{u}$ and $\mathbf{v} \in \mathcal{E}$), and observing that for all $\boldsymbol{h} \in \mathcal{H}$, there exists $\boldsymbol{\alpha} \in \mathbb{C}^N$ such that $\boldsymbol{h} = \mathbf{K}\boldsymbol{\alpha}$, we in particular have

$$\langle \boldsymbol{h} \mid \mathbf{K}\mathbf{v}\rangle_{\mathcal{H}} = \langle \boldsymbol{h} \mid \mathbf{v}\rangle_{\mathcal{E}}, \ \boldsymbol{h} \in \mathcal{H} \text{ and } \mathbf{v} \in \mathcal{E}. \tag{2.3}$$

The matrix $\mathbf{K}$ can in particular be regarded as an operator from, and to, $\mathcal{E}$ or $\mathcal{H}$. In (2.2), the trace norm then corresponds to the squared HS norm of the PSD matrix $\mathbf{K} - \hat{\mathbf{K}}(\mathbb{I})$ when interpreted as an operator from $\mathcal{E}$ to $\mathcal{H}$; indeed, setting $P_{0\mathbb{I}} = \mathbf{I} - P_{\mathbb{I}}$ (so that $\mathbf{K} - \hat{\mathbf{K}}(\mathbb{I}) = P_{0\mathbb{I}}\mathbf{K} = \mathbf{K}P_{0\mathbb{I}}^*$) and observing that the matrix $P_{0\mathbb{I}}$ is an orthogonal projection on $\mathcal{H}$, from (2.3), we obtain (see also Appendix A)

$$\sum_{i\in[N]} \|P_{0\mathbb{I}}\mathbf{K}\mathbf{e}_i\|_{\mathcal{H}}^2 = \sum_{i\in[N]} \langle P_{0\mathbb{I}}\mathbf{K}\mathbf{e}_i \mid \mathbf{K}\mathbf{e}_i\rangle_{\mathcal{H}} = \sum_{i\in[N]} \langle P_{0\mathbb{I}}\mathbf{K}\mathbf{e}_i \mid \mathbf{e}_i\rangle_{\mathcal{E}} = \operatorname{trace}(P_{0\mathbb{I}}\mathbf{K}).$$

Also, the Frobenius and spectral norms correspond to the HS and spectral norms, respectively, of the matrix $\mathbf{K} - \hat{\mathbf{K}}(\mathbb{I})$ when regarded as an operator on $\mathcal{E}$. ◁

2.2. First relaxation: selection vectors. For a *selection vector* $\boldsymbol{\upsilon} = (\upsilon_i)_{i\in[N]} \in \mathbb{R}^N$, we set $\mathbb{I}_{\boldsymbol{\upsilon}} = \{i \in [N] | \upsilon_i \neq 0\}$ and we refer to $\mathbb{I}_{\boldsymbol{\upsilon}}$ as the *support* of $\boldsymbol{\upsilon}$. Through its support, the vector $\boldsymbol{\upsilon}$ characterises a subset of columns of $\mathbf{K}$; following Remark 2.1, we introduce the simplified notations

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

We then define the error maps (notice the square in the definition of C_{F} and C_{sp})

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

Theorem 2.2. *The error maps C_{X}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$, are convex on the convex cone $\mathbb{R}^N_{\geqslant 0}$, and for $\boldsymbol{\upsilon} \in \mathbb{R}^N_{\geqslant 0}$ and $\boldsymbol{\eta} \in \mathbb{R}^N_{\geqslant 0}$, we have $\lim_{\rho\to 0^+} \frac{1}{\rho}\big[C_{\mathrm{X}}\big(\boldsymbol{\upsilon}+\rho(\boldsymbol{\eta}-\boldsymbol{\upsilon})\big) - C_{\mathrm{X}}(\boldsymbol{\upsilon})\big] \in \{-\infty, 0\}$, that is, the directional derivatives of these maps take values in the discrete set $\{-\infty, 0\}$.*

Theorem 2.2 illustrates that the error maps induced by the trace, Frobenius and spectral norms are akin to convex piecewise-constant functions on $\mathbb{R}^N_{\geqslant 0}$ (for the trace and Frobenius norms, an equivalent of this result can be found in [8]); see Figure 1 for an illustration. The selection-vector formulation can hence be regarded as a *non-differentiable convex relaxation* of the CSP. Introducing $|\boldsymbol{\upsilon}| = (|\upsilon_i|)_{i\in[N]} \in \mathbb{R}^N_{\geqslant 0}$, we may observe that $C_{\mathrm{X}}(\boldsymbol{\upsilon}) = C_{\mathrm{X}}(|\boldsymbol{\upsilon}|)$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$.

2.3. Second relaxation: quadrature approximation. Following Remark 2.1, we introduce

$$\mathrm{HS}(\mathcal{H}) = \{\mathbf{M} \in \mathbb{C}^{N\times N} \,|\, \operatorname{span}\{\mathbf{M}\mathbf{K}\} \subseteq \mathcal{H}\},$$

that is, a matrix $\mathbf{M}$ belongs to $\mathrm{HS}(\mathcal{H})$ if and only if $\mathbf{M}\boldsymbol{h} \in \mathcal{H}$ for all $\boldsymbol{h} \in \mathcal{H}$. Observing that for any orthonormal basis (ONB) $\{\boldsymbol{h}_j\}_{j\in\mathbb{J}}$ of $\mathcal{H}$, $\mathbb{J} \subseteq [N]$, we have $\mathbf{K} = \sum_{j\in\mathbb{J}} \boldsymbol{h}_j \boldsymbol{h}_j^*$ (see e.g. [19]), we set

$$\langle \mathbf{M} \,|\, \mathbf{T} \rangle_{\mathrm{HS}(\mathcal{H})} = \sum_{j\in\mathbb{J}} \langle \mathbf{M}\boldsymbol{h}_j \,|\, \mathbf{T}\boldsymbol{h}_j \rangle_{\mathcal{H}} = \mathrm{trace}(\mathbf{K}\mathbf{M}^*\mathbf{K}^\dagger\mathbf{T}), \ \mathbf{M} \text{ and } \mathbf{T} \in \mathrm{HS}(\mathcal{H}). \tag{2.4}$$

Endowed with the Hermitian form $\langle \cdot \,|\, \cdot \rangle_{\mathrm{HS}(\mathcal{H})}$, the linear space $\mathrm{HS}(\mathcal{H})$ is a semi-Hilbert space, and $\|\mathbf{M}\|_{\mathrm{HS}(\mathcal{H})} = 0$ if and only if $\mathbf{M}\mathbf{K} = 0$ (see Remark 2.3). If $\mathbf{K}$ is invertible, $\mathrm{HS}(\mathcal{H})$ is a Hilbert space.

Remark 2.3. When the matrix $\mathbf{K}$ is singular, the matrices representing a given operator on $\mathcal{H}$ are non-unique. Indeed, for $\mathbf{v} \in \mathbb{C}^N$, with $\mathbf{v} \neq 0$ such that $\mathbf{K}\mathbf{v} = 0$, we have $\mathbf{v}^*\boldsymbol{h} = 0$, $\boldsymbol{h} \in \mathcal{H}$; for $\mathbf{M} \in \mathrm{HS}(\mathcal{H})$ and $\mathbf{u} \in \mathbb{C}^N$, we obtain $(\mathbf{M} + \mathbf{u}\mathbf{v}^*)\boldsymbol{h} = \mathbf{M}\boldsymbol{h}$, so that the matrices $\mathbf{M}$ and $\mathbf{M} + \mathbf{u}\mathbf{v}^*$ represent the same operator on $\mathcal{H}$. ◁

A selection vector $\boldsymbol{\upsilon} \in \mathbb{R}^N$ can be regarded as a signed measure on $[N]$, and as such, defines together with $\mathbf{K}$ a discrete integral operator of the form $\mathbf{u} \mapsto \mathbf{K}\mathbf{V}\mathbf{u}$, $\mathbf{u} \in \mathbb{C}^N$, with $\mathbf{V} = \mathrm{diag}(\boldsymbol{\upsilon}) \in \mathbb{C}^{N\times N}$ the diagonal matrix with diagonal $\boldsymbol{\upsilon}$. The matrix $\mathbf{K}\mathbf{V}$ belongs to $\mathrm{HS}(\mathcal{H})$ and $\mathrm{span}\{\mathbf{K}\mathbf{V}\} = \mathcal{H}_{\boldsymbol{\upsilon}}$ (so that the matrices $\mathbf{K}\mathbf{V}$ and $\hat{\mathbf{K}}(\boldsymbol{\upsilon})$ relate to the same subset of columns of $\mathbf{K}$). Let $\boldsymbol{\omega} \in \mathbb{R}^N$ be another selection vector, and set $\mathbf{W} = \mathrm{diag}(\boldsymbol{\omega})$. From (2.4), we have

$$\langle \mathbf{K}\mathbf{W} \,|\, \mathbf{K}\mathbf{V} \rangle_{\mathrm{HS}(\mathcal{H})} = \mathrm{trace}(\mathbf{K}\mathbf{W}\mathbf{K}\mathbf{K}^\dagger\mathbf{K}\mathbf{V}) = \mathrm{trace}(\mathbf{K}\mathbf{W}\mathbf{K}\mathbf{V}) = \boldsymbol{\omega}^*\mathbf{S}\boldsymbol{\upsilon}, \tag{2.5}$$

where $\mathbf{S} = \overline{\mathbf{K}} \odot \mathbf{K}$ (element-wise product) is the $N \times N$ PSD matrix with i, j entry $|\mathbf{K}_{i,j}|^2$, the squared modulus of the i, j entry of $\mathbf{K}$ (the matrix $\mathbf{S}$ is real symmetric). Introducing $\mathbb{1} = (1)_{i\in[N]} \in \mathbb{R}^N$, we in particular have $\mathrm{diag}(\mathbb{1}) = \mathbf{I}$ (the identity matrix) and $\|\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{H})} = \mathbb{1}^*\mathbf{S}\mathbb{1} = \|\mathbf{K}\|^2_{\mathrm{F}}$.

We denote by $D : \mathbb{R}^N \to \mathbb{R}_{\geqslant 0}$ the error map defined as

$$D(\boldsymbol{\upsilon}) = \|\mathbf{K} - \mathbf{K}\mathbf{V}\|^2_{\mathrm{HS}(\mathcal{H})} = (\mathbb{1} - \boldsymbol{\upsilon})^*\mathbf{S}(\mathbb{1} - \boldsymbol{\upsilon}) = \|\mathbf{K}\|^2_{\mathrm{F}} + \boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon} - 2\boldsymbol{\mathfrak{g}}^*\boldsymbol{\upsilon}, \boldsymbol{\upsilon} \in \mathbb{R}^N, \tag{2.6}$$

with $\boldsymbol{\mathfrak{g}} = \mathbf{S}\mathbb{1} \in \mathbb{R}^N_{\geqslant 0}$ (we refer to $\boldsymbol{\mathfrak{g}}$ as the *target potential*; see Remark 2.4). The error map D is convex on $\mathbb{R}^N$, and the gradient of D at $\boldsymbol{\upsilon}$ is $\nabla D(\boldsymbol{\upsilon}) = 2(\mathbf{S}\boldsymbol{\upsilon} - \boldsymbol{\mathfrak{g}})$. The relation between D and the error maps C_{tr}, C_{F} and C_{sp} is further discussed in Section 2.5 (see also [8]).

Remark 2.4. Following Remark 2.1, 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 $\langle \boldsymbol{g} \,|\, \boldsymbol{j} \rangle_{\mathcal{G}} = \boldsymbol{g}^*\mathbf{S}^\dagger\boldsymbol{j}$, $\boldsymbol{g}$ and $\boldsymbol{j} \in \mathcal{G}$. In view of (2.5), we have

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

We refer to $\mathbf{S}\boldsymbol{\upsilon}$ as the *potential* of $\boldsymbol{\upsilon}$ in $\mathcal{G}$, and to $\|\mathbf{S}\boldsymbol{\upsilon}\|^2_{\mathcal{G}} = \|\mathbf{K}\mathbf{V}\|^2_{\mathrm{HS}(\mathcal{H})} = \boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon}$ as the *energy* of $\boldsymbol{\upsilon}$ with respect to $\mathbf{S}$. The map $(\boldsymbol{\omega}, \boldsymbol{\upsilon}) \mapsto \|\mathbf{K}\mathbf{W} - \mathbf{K}\mathbf{V}\|_{\mathrm{HS}(\mathcal{H})}$ can then be interpreted as a generalised *integral probability metric*, or *maximum mean discrepancy* (see e.g. [22, 16, 8]). ◁

2.4. Invariance under rescaling. For $\boldsymbol{\upsilon} \in \mathbb{R}^N$ and $c > 0$, we have $\mathbb{I}_{\boldsymbol{\upsilon}} = \mathbb{I}_{c\boldsymbol{\upsilon}}$; the error maps C_{X}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$, are thus invariant under rescaling, that is, $C_{\mathrm{X}}(c\boldsymbol{\upsilon}) = C_{\mathrm{X}}(\boldsymbol{\upsilon})$. To enforce a similar invariance within (2.6), we introduce the error map

$$R(\boldsymbol{\upsilon}) = \min_{c\geqslant 0} D(c\boldsymbol{\upsilon}) = \begin{cases} \|\mathbf{K}\|^2_{\mathrm{F}} - (\boldsymbol{\mathfrak{g}}^*\boldsymbol{\upsilon})^2/(\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon}) \text{ if } \boldsymbol{\mathfrak{g}}^*\boldsymbol{\upsilon} > 0, \\ \|\mathbf{K}\|^2_{\mathrm{F}} \text{ otherwise,} \end{cases} \tag{2.7}$$

and we set $\mathscr{D} = \{\boldsymbol{v} \in \mathbb{R}^N | \mathfrak{g}^* \boldsymbol{v} > 0\}$. From the Cauchy-Schwarz (CS) inequality, if $\boldsymbol{v} \in \mathscr{D}$, then $\boldsymbol{v}^* \mathbf{S} \boldsymbol{v} > 0$; we indeed have $|\mathfrak{g}^* \boldsymbol{v}|^2 = |\mathbb{1}^* \mathbf{S} \boldsymbol{v}|^2 \leqslant (\mathbb{1}^* \mathbf{S} \mathbb{1})(\boldsymbol{v}^* \mathbf{S} \boldsymbol{v})$. We also have $R(\boldsymbol{v}) = D(c_{\boldsymbol{v}} \boldsymbol{v})$, with

$$c_{\boldsymbol{v}} = \begin{cases} (\mathfrak{g}^* \boldsymbol{v})/(\boldsymbol{v}^* \mathbf{S} \boldsymbol{v}) \text{ if } \boldsymbol{v} \in \mathscr{D}, \\ 0 \text{ otherwise.} \end{cases}$$

The appearance of the error maps D, R and C_{F} over $\mathbb{R}^N_{\geqslant 0}$ is illustrated in Figure 1.

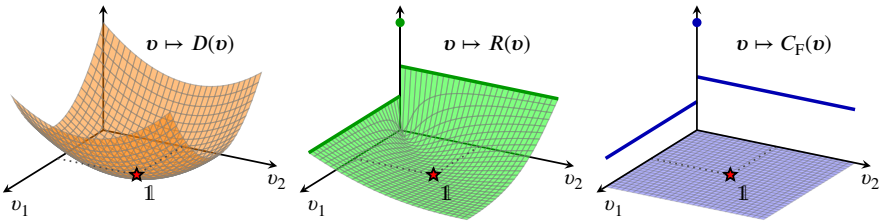

Figure 1. *Schematic representation of the error maps D, R and C_{F} on $\mathbb{R}^N_{\geqslant 0}$; the red star represents the selection vector $\mathbb{1} \in \mathbb{R}^N$. The presented graphs correspond to a 2×2 matrix $\mathbf{K}$ such that $\mathbf{K}_{1,1} = 1.225$, $\mathbf{K}_{2,2} = 0.894$ 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 these maps along the horizontal axes.*

For $\boldsymbol{\eta} \in \mathbb{R}^N$, the directional derivative $\Theta(\boldsymbol{v}; \boldsymbol{\eta})$ of R at $\boldsymbol{v} \in \mathbb{R}^N$ along $\boldsymbol{\eta} - \boldsymbol{v}$ is given by

$$\Theta(\boldsymbol{v}; \boldsymbol{\eta}) = \lim_{\rho \to 0^+} \frac{1}{\rho} \big[R\big(\boldsymbol{v} + \rho(\boldsymbol{\eta} - \boldsymbol{v})\big) - R(\boldsymbol{v}) \big] = \begin{cases} -\infty \text{ if } \boldsymbol{v} \in \mathscr{Z} \text{ and } \boldsymbol{\eta} \in \mathscr{D}, \\ 2c_{\boldsymbol{v}} (\boldsymbol{\eta} - \boldsymbol{v})^* (c_{\boldsymbol{v}} \mathbf{S} \boldsymbol{v} - \mathfrak{g}) \text{ otherwise,} \end{cases} \tag{2.8}$$

with $\mathscr{Z} = \{\boldsymbol{v} \in \mathbb{R}^N | \mathbf{S} \boldsymbol{v} = 0\}$. As $\mathscr{D} \cap \mathscr{Z} = \emptyset$, the gradient of R at $\boldsymbol{v} \in \mathscr{D}$ is $\nabla R(\boldsymbol{v}) = 2c_{\boldsymbol{v}}(c_{\boldsymbol{v}} \mathbf{S} \boldsymbol{v} - \mathfrak{g})$. We may observe that for $\boldsymbol{v} \in \mathscr{D}$, $\boldsymbol{v}^*(c_{\boldsymbol{v}} \mathbf{S} \boldsymbol{v} - \mathfrak{g}) = 0$.

Theorem 2.5. *The map R is quasiconvex on $\mathbb{R}^N$, and pseudoconvex on the convex cone $\mathscr{D}$.*

For $\boldsymbol{v}^\star = c\mathbb{1} + \boldsymbol{\epsilon}$, with $c > 0$ and $\boldsymbol{\epsilon} \in \mathbb{R}^N$ such that $\mathbf{S}\boldsymbol{\epsilon} = 0$, we have $R(\boldsymbol{v}^\star) = 0$, and R is thus minimum at $\boldsymbol{v}^\star$. For suitable step sizes, the pseudoconvexity of R on $\mathscr{D}$ ensures the convergence to such a minimum of any gradient descent starting from a vector in $\mathscr{D}$. Lemma 2.6 provides an analytical expression for the optimal step size and for the improvement induced by a descent with optimal step size in the framework of interest for Section 3.

Lemma 2.6. *For $\boldsymbol{v} \in \mathscr{D}$ and $\boldsymbol{\eta} \in \mathbb{R}^N$ such that $\Theta(\boldsymbol{v}; \boldsymbol{\eta}) < 0$ and $\Theta(\boldsymbol{\eta}; \boldsymbol{v}) \leqslant 0$, 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*

$$r = \frac{T_1}{T_1 + T_2},\ T_1 = (\boldsymbol{v}^* \mathbf{S} \boldsymbol{v})(\mathfrak{g}^* \boldsymbol{\eta}) - (\mathfrak{g}^* \boldsymbol{v})(\boldsymbol{v}^* \mathbf{S} \boldsymbol{\eta}) \text{ and } T_2 = (\boldsymbol{\eta}^* \mathbf{S} \boldsymbol{\eta})(\mathfrak{g}^* \boldsymbol{v}) - (\mathfrak{g}^* \boldsymbol{\eta})(\boldsymbol{v}^* \mathbf{S} \boldsymbol{\eta}); \tag{2.9}$$

introducing $\mathcal{I}(\boldsymbol{v}; \boldsymbol{\eta}) = R(\boldsymbol{v}) - R\big(\boldsymbol{v} + r(\boldsymbol{\eta} - \boldsymbol{v})\big) \geqslant 0$, we then have

$$\mathcal{I}(\boldsymbol{v}; \boldsymbol{\eta}) = \big(\boldsymbol{\eta}^* (c_{\boldsymbol{v}} \mathbf{S} \boldsymbol{v} - \mathfrak{g})\big)^2 \big/ \big((\boldsymbol{\eta}^* \mathbf{S} \boldsymbol{\eta}) - (\boldsymbol{v}^* \mathbf{S} \boldsymbol{\eta})^2 / (\boldsymbol{v}^* \mathbf{S} \boldsymbol{v})\big). \tag{2.10}$$

2.5. Additional error maps and further properties. In $\mathrm{HS}(\mathcal{H})$, the Nyström approximation of $\mathbf{K}$ relates to the approximation of the underlying operator on $\mathcal{H}$ through projections. For $\boldsymbol{\upsilon} \in \mathbb{R}^N$, we have (see Lemma A.1 and (B.2), in appendix)

$$\|\mathbf{K} - P_{\boldsymbol{\upsilon}}\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{H})} = \langle \mathbf{K} - \hat{\mathbf{K}}(\boldsymbol{\upsilon}) \,|\, \mathbf{K}\rangle_{\mathrm{F}} \quad \text{and} \quad \|\mathbf{K} - P_{\boldsymbol{\upsilon}}\mathbf{K}P_{\boldsymbol{\upsilon}}\|^2_{\mathrm{HS}(\mathcal{H})} = \|\mathbf{K}\|^2_{\mathrm{F}} - \|\hat{\mathbf{K}}(\boldsymbol{\upsilon})\|^2_{\mathrm{F}},$$

with $\langle \cdot \,|\, \cdot \rangle_{\mathrm{F}}$ the Frobenius inner product on $\mathbb{C}^{N\times N}$. This observation suggests the definition of the additional error maps

$$C_{\mathrm{P}}(\boldsymbol{\upsilon}) = \langle \mathbf{K} - \hat{\mathbf{K}}(\boldsymbol{\upsilon}) \,|\, \mathbf{K}\rangle_{\mathrm{F}} \quad \text{and} \quad C_{\mathrm{PP}}(\boldsymbol{\upsilon}) = \|\mathbf{K}\|^2_{\mathrm{F}} - \|\hat{\mathbf{K}}(\boldsymbol{\upsilon})\|^2_{\mathrm{F}};$$

these maps are of the same type as the maps C_{X}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$, as illustrated by Proposition 2.7.

Proposition 2.7. *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 $\{-\infty, 0\}$.*

The following Lemma 2.8 shows that the error maps C_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{sp}, \mathrm{P}, \mathrm{PP}\}$, are upper-bounded by R. We may also observe that

$$C_{\mathrm{X}}(\mathbb{1}) = R(\mathbb{1}) = 0, \mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}, \mathrm{P}, \mathrm{PP}\} \quad \text{and} \quad C_{\mathrm{X}}(0) = R(0) = \|\mathbf{K}\|^2_{\mathrm{F}}, \mathrm{X} \in \{\mathrm{F}, \mathrm{P}, \mathrm{PP}\}.$$

Lemma 2.8. *For all $\boldsymbol{\upsilon} \in \mathbb{R}^N$, we have $C_{\mathrm{sp}}(\boldsymbol{\upsilon}) \leqslant C_{\mathrm{F}}(\boldsymbol{\upsilon}) \leqslant C_{\mathrm{P}}(\boldsymbol{\upsilon}) \leqslant C_{\mathrm{PP}}(\boldsymbol{\upsilon}) \leqslant R(\boldsymbol{\upsilon}) \leqslant D(\boldsymbol{\upsilon})$; in addition, $C_{\mathrm{PP}}(\mathbf{e}_i) = R(\mathbf{e}_i)$, $i \in [N]$.*

In view of the above, the error map R can be regarded as a differentiable surrogate for the characterisation of samples of columns for Nyström approximation (see also [9, 11, 8]). In the forthcoming Section 3, we describe a class of sequential sampling strategies driven by the gradient of R.

3. Gradient-based sequential sampling. From now on, we assume that the diagonal entries of $\mathbf{K}$ are strictly positive, so that $\mathbb{R}^N_{\geqslant 0}\backslash\{0\} \subset \mathscr{D}$ (this assumption is not restrictive: if a diagonal entry of $\mathbf{K}$ is equal to 0, then by CS, the corresponding row and column of $\mathbf{K}$ are zero vectors). For $\mathbf{f} = (f_i)_{i\in[N]} \in \mathbb{R}^N_{>0}$ and $\varkappa > 0$, we introduce

$$\mathcal{A}_{\mathbf{f}} = \{\boldsymbol{\upsilon} \in \mathbb{R}^N_{\geqslant 0} | \mathbf{f}^*\boldsymbol{\upsilon} = \varkappa\} \subset \mathscr{D};$$

we refer to $\mathbf{f}$ as the *restriction vector*. The set $\mathcal{A}_{\mathbf{f}}$ is convex, and its extreme points are the vectors $\{\boldsymbol{\xi}_i\}_{i\in[N]}$, with $\boldsymbol{\xi}_i = \varkappa \mathbf{e}_i / f_i \in \mathbb{R}^N_{\geqslant 0}$. Below, we describe a column-sampling procedure based on the minimisation of R over $\mathcal{A}_{\mathbf{f}}$ via line search with sparse initialisation and sparse descent directions (specifically, the directions defined by the extreme points of $\mathcal{A}_{\mathbf{f}}$); sparsity of the samples is enforced via early stopping. Many variants may be considered (see for instance Remarks 3.1, 3.2 and 3.3), and stochastic variants are discussed in Section 4. Due to the invariance under rescaling of R, the value of $\varkappa$ does not impact the sampling procedure (and we may thus set $\varkappa = 1$, for instance).

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

$$b \in \arg\min_{i\in[N]} R(\boldsymbol{\xi}_i) = \arg\max_{i\in[N]} \mathfrak{g}_i^2 / \mathbf{S}_{i,i}, \text{ with } \mathfrak{g}_i = \mathbf{e}_i^*\mathfrak{g} \text{ the } i\text{-th entry of } \mathfrak{g} = \mathbf{S}\mathbb{1}, \tag{3.1}$$

and the selection vector at step $q \in \mathbb{N}$ is denoted by $\boldsymbol{\upsilon}^{(q)}$. An iteration of our sampling procedure consists of selecting a direction $\boldsymbol{\xi}_i - \boldsymbol{\upsilon}^{(q)}$, with $i \in [N]$ such that $\Theta(\boldsymbol{\upsilon}^{(q)}; \boldsymbol{\xi}_i) < 0$, and of next performing a

descent with the corresponding optimal step size $r^{(q)}$ given by (2.9). As descent direction, we consider the Frank-Wolfe (FW) direction $\boldsymbol{\xi}_u - \boldsymbol{\upsilon}^{(q)}$, with

$$u \in \arg\min_{i\in[N]} \Theta(\boldsymbol{\upsilon}^{(q)}; \boldsymbol{\xi}_i) = \arg\min_{i\in[N]} [\nabla R(\boldsymbol{\upsilon}^{(q)})]_i / f_i. \tag{3.2}$$

Notably, the initialisation of the procedure via (3.1) ensures that if $\Theta(\boldsymbol{\upsilon}^{(q)}; \boldsymbol{\xi}_i) < 0$, $i \in [N]$, then $\Theta(\boldsymbol{\xi}_i; \boldsymbol{\upsilon}^{(q)}) < 0$ (by pseudoconvexity, we would otherwise have $R(\boldsymbol{\xi}_i) < R(\boldsymbol{\xi}_b)$, which is impossible by definition of $\boldsymbol{\xi}_b$). The descents therefore necessarily occur within the framework of Lemma 2.6, and we always have $\boldsymbol{\upsilon}^{(q)} \in \mathcal{A}_{\mathbf{f}}$ (indeed, $\mathcal{A}_{\mathbf{f}}$ is convex and $r^{(q)} \in [0, 1]$).

A pseudocode of the procedure is given in Algorithm 1. The algorithm produces a sequence $\boldsymbol{\upsilon}^{(1)}, \boldsymbol{\upsilon}^{(2)}, \cdots$ of selection vectors with increasing support. At stage $q \in \mathbb{N}$, the number m_q of non-zero entries of $\boldsymbol{\upsilon}^{(q)}$ verifies $m_q \leqslant \min(q, N)$ (see also Remark 3.2). The algorithm stops when $q = Q$, where $Q \in \mathbb{N}$ is a given maximum number of iterations, with in practice $Q \ll N$ (different stopping rules could be considered, based on m_q for instance); the algorithm also stops if $\boldsymbol{\upsilon}^{(q)}$ minimises R over $\mathcal{A}_{\mathbf{f}}$ (that is, if there are no descent directions; this situation is unlikely, especially for $q \ll N$). We may observe that $\boldsymbol{\upsilon}^\star = \varkappa \mathbb{1}/(\mathbf{f}^* \mathbb{1}) \in \mathcal{A}_{\mathbf{f}}$ verifies $R(\boldsymbol{\upsilon}^\star) = 0$.

Algorithm 1 Column sampling with FW direction and optimal step size.

Input: matrix $\mathbf{S}$; vector $\mathbf{f}$; number of iterations $Q \in \mathbb{N}$;
  Preliminary: compute $\mathfrak{g} = \mathbf{S}\mathbb{1}$ (stochastic approximations may be considered, see Section 4);
  Initialisation: compute $b \in [N]$ using (3.1); set $q = 1$, $\boldsymbol{\upsilon}^{(1)} = \boldsymbol{\xi}_b$ and $\mathbb{I}_{\boldsymbol{\upsilon}^{(1)}} = \{b\}$;
  while $q < Q$ and $R(\boldsymbol{\upsilon}^{(q)}) > 0$ **do**
    compute $u \in [N]$ using (3.2);
    compute the optimal step size $r^{(q)}$ from (2.9) with $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}^{(q)}$ and $\boldsymbol{\eta} = \boldsymbol{\xi}_u$;
    set $\boldsymbol{\upsilon}^{(q+1)} = (1 - r^{(q)})\boldsymbol{\upsilon}^{(q)} + r^{(q)}\boldsymbol{\xi}_u$ and $\mathbb{I}_{\boldsymbol{\upsilon}^{(q+1)}} = I_{\boldsymbol{\upsilon}^{(q)}} \cup \{u\}$; increment q;
  end while
Output: subset $\mathbb{I}_{\boldsymbol{\upsilon}^{(q)}} \subseteq [N]$;

The implementation of Algorithm 1 involves the preliminary computation of the target potential $\mathfrak{g} = \mathbf{S}\mathbb{1}$. Although easily parallelisable, this operation has a $\mathcal{O}(N^2)$ worst-case time complexity (it requires reading every entry of $\mathbf{S}$ once); this cost can nevertheless be reduced by considering stochastic approximations of $\mathfrak{g}$, as discussed in Section 4. Once $\mathfrak{g}$ is known, each iteration of Algorithm 1 has a $\mathcal{O}(N)$ time complexity. Notably, for $q \in \mathbb{N}$, we for instance have

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

so that sparse updates of the terms $\mathbf{S}\boldsymbol{\upsilon}$, $\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon}$ and $\mathfrak{g}^*\boldsymbol{\upsilon}$ can be easily implemented.

In view of (3.2), the sequence of subsets $\mathbb{I}_{\boldsymbol{\upsilon}^{(1)}} \subseteq \mathbb{I}_{\boldsymbol{\upsilon}^{(2)}} \subseteq \cdots$ generated by Algorithm 1 depends 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 to be a relevant choice. Variants of Algorithm 1 producing sequences of subsets that are independent of the choice of $\mathbf{f}$ are discussed in Remark 3.4.

Remark 3.1 (Best-improvement direction). Instead of considering the steepest conditional descent directions (3.2), we may combine the information provided by (2.8) and (2.10) to characterise the

conditional descent directions inducing the best one-step-ahead improvements. In Algorithm 1, we may hence replace the FW direction (3.2) by the *best-improvement* (BI) direction

$$u \in \arg\max_{i\in[N]} \{\mathcal{I}(\boldsymbol{\upsilon}^{(q)};\boldsymbol{\xi}_i) | \Theta(\boldsymbol{\upsilon}^{(q)};\boldsymbol{\xi}_i) < 0\}.$$

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. ◁

Remark 3.2 (Enforcing the selection of new columns). In Algorithm 1, at step $q \in \mathbb{N}$, the FW direction (3.2) might lead to the selection of a column which already belongs to the sample, that is, we may have $u \in \mathbb{I}_{\boldsymbol{\upsilon}^{(q)}}$; we refer to such an event as a *correction step* (a similar observation holds for the BI variant of the algorithm). To enforce the selection of a new column at each iteration, we may replace the FW direction (3.2) by

$$u \in \arg\min_{i\in[N]} \{\Theta(\boldsymbol{\upsilon}^{(q)};\boldsymbol{\xi}_i) | i \notin \mathbb{I}_{\boldsymbol{\upsilon}^{(q)}} \text{ and } \Theta(\boldsymbol{\upsilon}^{(q)};\boldsymbol{\xi}_i) < 0\}; \tag{3.3}$$

if the set characterising (3.3) is empty, the sampling should stop (or an alternative direction should be considered). Such a variant of Algorithm 1 ensures a faster, although less accurate, exploration of the columns of $\mathbf{K}$; it appears to be of particular interest in the stochastic setting of Section 4. ◁

Remark 3.3 (Weight optimisation). For a subset $\mathbb{I} \subseteq [N]$ of size $m \in$, let $\tilde{\boldsymbol{\upsilon}}(\mathbb{I}) \in \mathbb{R}^N_{\geqslant 0}$ be a vector minimising D over the set of all selection vectors $\boldsymbol{\upsilon} \in \mathbb{R}^N_{\geqslant 0}$ such that $\mathbb{I}_{\boldsymbol{\upsilon}} \subseteq \mathbb{I}$ (the entries of the PSD matrix $\mathbf{S}$ being non-negative, such a vector always exists). The non-trivial entries $[\tilde{\boldsymbol{\upsilon}}(\mathbb{I})]_{\mathbb{I}}$ of $\tilde{\boldsymbol{\upsilon}}(\mathbb{I})$ are provided by solutions to the quadratic program (QP) associated with the minimisation of the function $\boldsymbol{x} \mapsto \boldsymbol{x}^* \mathbf{S}_{\mathbb{I},\mathbb{I}} \boldsymbol{x} - 2\mathbf{g}^*_{\mathbb{I}} \boldsymbol{x}$ over $\mathbb{R}^m_{\geqslant 0}$. The rescaled vector $\boldsymbol{\upsilon}(\mathbb{I}) = \varkappa \tilde{\boldsymbol{\upsilon}}(\mathbb{I})/(\mathbf{f}^* \tilde{\boldsymbol{\upsilon}}(\mathbb{I})) \in \mathcal{A}_{\mathbf{f}}$ then minimises R over the set of all selection vectors $\boldsymbol{\upsilon} \in \mathcal{A}_{\mathbf{f}}$ such that $\mathbb{I}_{\boldsymbol{\upsilon}} \subseteq \mathbb{I}$. In Algorithm 1 and its BI variant, at iteration $q \in \mathbb{N}$, rather than performing a descent with optimal step size, we may instead set $\boldsymbol{\upsilon}^{(q+1)} = \boldsymbol{\upsilon}(\mathbb{I}_{\boldsymbol{\upsilon}^{(q)}} \cup \{u\})$. We refer to this modified update rule as *weight optimisation* (WO); the algorithm then converges in at most N iterations. In terms of numerical complexity, and in comparison to descents with optimal step sizes, for the WO variants, the computation of $\boldsymbol{\upsilon}^{(q+1)}$ involves solving a QP over $\mathbb{R}^{m_q+1}$ (in practice, $\tilde{\boldsymbol{\upsilon}}^{(q)}$ may be used as a warm start for the computation of $\tilde{\boldsymbol{\upsilon}}^{(q+1)}$). As a technical remark, for $q \in \mathbb{N}$, the support of $\boldsymbol{\upsilon}^{(q+1)}$ might sometimes be a strict subset of $\mathbb{I}_{\boldsymbol{\upsilon}^{(q)}} \cup \{u\}$; this situation occurs when some entries of the solution to the underlying QP are zero. In the experiments of Section 5, instead of the true support $\mathbb{I}_{\boldsymbol{\upsilon}^{(q+1)}}$, we keep track of the *virtual support* $\tilde{\mathbb{I}}_{\boldsymbol{\upsilon}^{(q+1)}} = \tilde{\mathbb{I}}_{\boldsymbol{\upsilon}^{(q)}} \cup \{u\}$, so that $\mathrm{card}(\tilde{\mathbb{I}}_{\boldsymbol{\upsilon}^{(q)}}) = q$ for all $q \leqslant N$ (that is, once a column of $\mathbf{K}$ has been selected, it is kept inside the sample even if its associated weight vanishes at some stages of the optimisation process). ◁

Remark 3.4 (Restriction vector and BI direction). In Algorithm 1, the column of $\mathbf{K}$ selected at initialisation via (3.1) does not depend on the choice of $\mathbf{f}$. In the framework of Lemma 2.6, we in addition have $\nabla R(c\boldsymbol{\upsilon}) = \nabla R(\boldsymbol{\upsilon})/c$ and $\mathcal{I}(\boldsymbol{\upsilon};\boldsymbol{\eta}) = \mathcal{I}(c\boldsymbol{\upsilon};\tilde{c}\boldsymbol{\eta})$, c and $\tilde{c} > 0$. Consequently, the sequences of column subsets produced by the BI and BI-WO variants of Algorithm 1 (that is, BI direction with optimal-step-size or WO update rule) do not depend on the choice of $\mathbf{f}$. For the optimal-step-size update rule, we should observe that a descent from $\boldsymbol{\upsilon}$ along $\boldsymbol{\eta} - \boldsymbol{\upsilon}$ and a descent from $c\boldsymbol{\upsilon}$ along $\tilde{c}\boldsymbol{\eta} - c\boldsymbol{\upsilon}$ lead to proportional selection vectors; a similar observation holds for the WO update rule. ◁

4. Stochastic approximation of the target potential. In practical applications, and due to its quadratic complexity in N, the preliminary computation of the target potential $\mathfrak{g}$ might be prohibitive. An alternative consists in relying on numerically affordable stochastic approximations of $\mathfrak{g}$. Many approaches may be considered, and below, we simply describe one possible way to proceed. We assume that $N > 1$.

Direct Monte Carlo approximation. The entries of $\mathfrak{g} = \mathbf{S}\mathbb{1}$ correspond to the row sums of $\mathbf{S}$; as such, they can be approximated by random sampling. The matrix $\mathbf{S}$ being PSD, we handle its diagonal separately and only sample off-diagonal entries of $\mathbf{S}$; each row is sampled independently of the others, with the same sample size $\ell \in \mathbb{N}$. The sampling is performed uniformly, and for simplicity, with replacement. For all $i \in [N]$, that is, for each row of $\mathbf{S}$, this operation amounts to forming a random multiset $\mathcal{S}_i$ of ℓ indices in $[N]\backslash\{i\}$. Denoting by $\mathbf{F}$ the $N \times N$ random matrix whose i, j entry counts the number of times $j \in [N]$ appears in $\mathcal{S}_i$ (so that $\mathbf{F}\mathbb{1} = \ell\mathbb{1}$), the random vector

$$\hat{\mathfrak{g}}_{\mathbf{F}} = \operatorname{diag}(\mathbf{S}) + \tfrac{(N-1)}{\ell}(\mathbf{S} \odot \mathbf{F})\mathbb{1}, \tag{4.1}$$

corresponds to an unbiased estimator of $\mathfrak{g}$. We may observe that the off-diagonal entries of $\mathbf{F}$ follow a binomial distribution with parameters ℓ and $\frac{1}{N-1}$.

Accounting for the symmetry of $\mathbf{S}$. In the framework of (4.1) and for ℓ fixed, the number of entries of $\mathbf{S}$ involved in the approximation of $\mathfrak{g}$ can be increased by accounting for the symmetry of $\mathbf{S}$. Indeed, if $i \in \mathcal{S}_j$, i and $j \in [N]$, $i \neq j$, that is, if $\mathbf{S}_{j,i}$ appears in the approximation of $\mathfrak{g}_j$ (the j-th entry of $\mathfrak{g}$), then $\mathbf{S}_{i,j} = \mathbf{S}_{j,i}$ may be incorporated into the approximation of $\mathfrak{g}_i$. The corresponding entries of $\mathbf{S}$ are provided by the matrix $\mathbf{F}^*$, and the random vector $\boldsymbol{l} = (l_i)_{i\in[N]} = \mathbf{F}^*\mathbb{1}$ indicates the number of additional entries per row of $\mathbf{S}$. The rows of $\mathbf{F}$ being independent random vectors, for all $i \in [N]$, the random variables $\{\mathbf{F}^*_{i,j}\}_{j\in[N]\backslash\{i\}}$ are independent, and l_i follows a binomial distribution with parameters $\ell(N-1)$ and $\frac{1}{N-1}$. Observing that $\mathbb{E}(\mathbf{F}^*_{i,j}|l_i) = \frac{l_i}{N-1}$ (conditional mean of $\mathbf{F}^*_{i,j}$ given l_i; see Lemma A.3), and denoting by $(\mathbb{1}/\boldsymbol{l}) = (1/\boldsymbol{l}) \in \mathbb{R}^N$ the vector with i-th entry $1/l_i$ if $l_i \neq 0$, and 0 otherwise (element-wise pseudoinversion), the random vector

$$\hat{\mathfrak{g}}_{\mathbf{F}^*} = \operatorname{diag}(\mathbf{S}) + \tfrac{(N-1)}{\boldsymbol{l}} \odot \big((\mathbf{S} \odot \mathbf{F}^*)\mathbb{1}\big)$$

is an unbiased estimator of $\mathfrak{g}$ (cf. *Bernoulli sampling*). From the independence between the rows of $\mathbf{F}$, for all $i \in [N]$, the i-th entries of $\hat{\mathfrak{g}}_{\mathbf{F}}$ and $\hat{\mathfrak{g}}_{\mathbf{F}^*}$ are independent; by considering sample-size-dependent convex combinations of these entries, we can form the unbiased estimator $\hat{\mathfrak{g}}_{\mathrm{sym}}$ of $\mathfrak{g}$, with

$$\hat{\mathfrak{g}}_{\mathrm{sym}} = \tfrac{\ell}{\ell+\boldsymbol{l}} \odot \hat{\mathfrak{g}}_{\mathbf{F}} + \tfrac{\boldsymbol{l}}{\ell+\boldsymbol{l}} \odot \hat{\mathfrak{g}}_{\mathbf{F}^*} = \operatorname{diag}(\mathbf{S}) + \tfrac{N-1}{\ell+\boldsymbol{l}} \odot \big([\mathbf{S} \odot (\mathbf{F} + \mathbf{F}^*)]\mathbb{1}\big),$$

where $\ell + \boldsymbol{l}$ is a simplified notation for $\ell\mathbb{1} + \boldsymbol{l}$. Accounting for the symmetry of $\mathbf{S}$ therefore results in increasing the number of independent samples per row of $\mathbf{S}$ at the cost of introducing a small residual dependence between the entries of $\hat{\mathfrak{g}}_{\mathrm{sym}}$ (indeed, contrary to $\hat{\mathfrak{g}}_{\mathbf{F}}$, the entries of $\hat{\mathfrak{g}}_{\mathbf{F}^*}$ are dependent); the mean of $\boldsymbol{l}$ being $\ell\mathbb{1}$, for each row, we in average double the sample size, hence reducing the variance of the approximation.

Remark 4.1. Computing a realisation of $\hat{\mathfrak{g}}_{\mathbf{F}}$, $\hat{\mathfrak{g}}_{\mathbf{F}^*}$ or $\hat{\mathfrak{g}}_{\mathrm{sym}}$ involves sampling $(\ell+1)N$ entries of $\mathbf{S}$. The time-complexity of forming such approximations is thus $\mathcal{O}(\ell N)$, with in practice $\ell \ll N$ (here, we assume that the complexity of the considered random generator does not depend on N); notably, if

ℓ is chosen independently of N (see Remark 4.2), the time complexity of forming an approximation is linear in N. The computation can in addition be easily parallelised. Observe that we should in practice not form the matrix $\mathbf{F}$, but instead simply sample a set of indices per row of $\mathbf{S}$ and use the corresponding entries of $\mathbf{S}$ to build the approximation. ◁

Sampling driven by an approximate potential. In (2.6) and (2.7), substituting $\mathfrak{g}$ with an approximation $\hat{\mathfrak{g}} \in \mathbb{R}^N_{\geqslant 0}\backslash\{0\}$ gives rise to approximate error maps $\hat{D}$ and $\hat{R}$. Let $\hat{\mathbb{1}} \in \mathbb{R}^N_{\geqslant 0}\backslash\{0\}$ be a vector minimising $\hat{D}$ over $\mathbb{R}^N_{\geqslant 0}$ (the non-negativity of the entries of the PSD matrix $\mathbf{S}$ ensures that such a vector always exists). When $\mathfrak{g}$ is replaced by $\hat{\mathfrak{g}}$, Algorithm 1 produces a sequence of selection vectors with increasing supports converging towards a vector minimising $\hat{R}$ over $\mathcal{A}_{\mathbf{f}}$, that is, a vector of the form $\varkappa\hat{\mathbb{1}}/(\mathbf{f}^*\hat{\mathbb{1}})$. A similar approximation scheme can be applied to the BI and WO variants of the algorithm. The same approximation of $\mathfrak{g}$ is used throughout the optimisation process (alternative strategies, where the approximation is updated during the optimisation process, could be considered).

Remark 4.2. When a realisation of $\hat{\mathfrak{g}}_{\mathbf{F}}$ or $\hat{\mathfrak{g}}_{\mathrm{sym}}$ is considered, for $\ell \ll N$, our experiments suggest that the underlying vector $\hat{\mathbb{1}} \in \mathbb{R}^N_{\geqslant 0}$ is often sparse (that is, $\hat{\mathbb{1}}$ has many zero entries); the sparsity of $\hat{\mathbb{1}}$ appears to decrease as ℓ increases. These observations suggest that the sample size ℓ should be selected in accordance with the number m of columns of $\mathbf{K}$ one wishes to extract; see Section 5 for illustrations. Following Remark 3.2, for the stochastic variant of Algorithm 1, we also observe that considering the modified FW direction (3.3) improves the behaviour of the sampling procedure by preventing the apparition of early correction steps resulting from the sparsity of $\hat{\mathbb{1}}$. Furthermore, in comparison to $\hat{\mathfrak{g}}_{\mathbf{F}}$, the reduced variance of the estimator $\hat{\mathfrak{g}}_{\mathrm{sym}}$ appears to have a beneficial impact on the column-sampling process. ◁

5. Experiments.

We now illustrate the behaviour of Algorithm 1 and its variants on a series of examples. To assess the accuracy of the Nyström approximation induced by a subset $\mathbb{I} \subseteq [N]$ of size $m \leqslant N$, we consider the *approximation factors* (see e.g. [5])

$$\mathscr{E}_{\mathrm{P}}(\mathbb{I}) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(\mathbb{I})\|_{\mathrm{HS}(\mathcal{H})}}{\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{HS}(\mathcal{H})}}, \mathscr{E}_{\mathrm{PP}}(\mathbb{I}) = \frac{\|\mathbf{K} - P_{\mathbb{I}}\mathbf{K}P_{\mathbb{I}}\|_{\mathrm{HS}(\mathcal{H})}}{\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{HS}(\mathcal{H})}} \text{ and } \mathscr{E}_{\mathrm{X}}(\mathbb{I}) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(\mathbb{I})\|_{\mathrm{X}}}{\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{X}}}, \tag{5.1}$$

$\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$, 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 5.1. 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); see Remarks 3.1 and 3.3. In the stochastic case, that is, when stochastic approximations of $\mathfrak{g}$ are considered (see Section 4), we rely on the estimator $\hat{\mathfrak{g}}_{\mathrm{sym}}$ and implement the modified FW direction (3.3); we refer to this variant as S-MFW. The affine restrictions are defined with $\mathbf{f} = \mathrm{diag}(\mathbf{K})$ and $\varkappa = 1$. Due to the specificity of our sampling procedures (which rely on early stopping of optimisation procedures 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.

We compare the resulting column samples with samples obtained through random sampling with respect to uniform weights and weights proportional to the squares of the diagonal entries of $\mathbf{K}$, *leverage-score-based* random sampling, and *determinantal-point-process-based* (DPP-based) random sampling; see for instance [7, 3, 12, 10, 14, 20, 6] for an overview.

5.1. 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 [15]). In this first experiment, we use the exact target potential $\mathfrak{g}$.

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 (these four error maps are considered since they take the same value at $\upsilon = 0$); in accordance with Lemma 2.8, the error maps C_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}, \mathrm{PP}\}$ are bounded by R. We observe a strong similarity between the evolution of these maps, further supporting the use of R as surrogate error map for Nyström approximation.

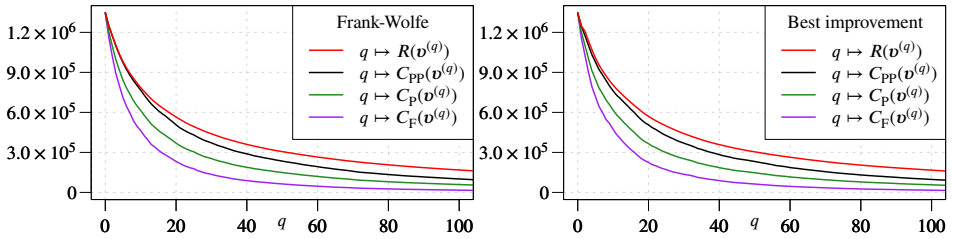

Figure 2. *For a random PSD matrix ($N = 1{,}500$), evolution of the value of the error maps R and C_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}, \mathrm{PP}\}$, during the 100 first iterations of Algorithm 1 (left) and its BI variant (right). The exact target potential $\mathfrak{g}$ is used. See Section 5.1 for more details.*

We then compare, for various sampling strategies, 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 results are presented in Figure 3. In the considered regime (that is, $m \ll N$), and for all the approximation factors, we observe that the Nyström approximations induced by Algorithm 1 and its variants are more accurate than the ones obtained using uniform random sampling, squared-diagonal random sampling or leverage-score-based random sampling. For this particular example, we may also notice the similarity and small variability of the approximation factors induced by the considered stochastic procedures.

5.2. Abalone data set. We consider the Abalone data set (see [18]). 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 eigenvalues of $\mathbf{K}$ exhibit relatively steep, moderate and shallow decays, respectively; see Figure 4.

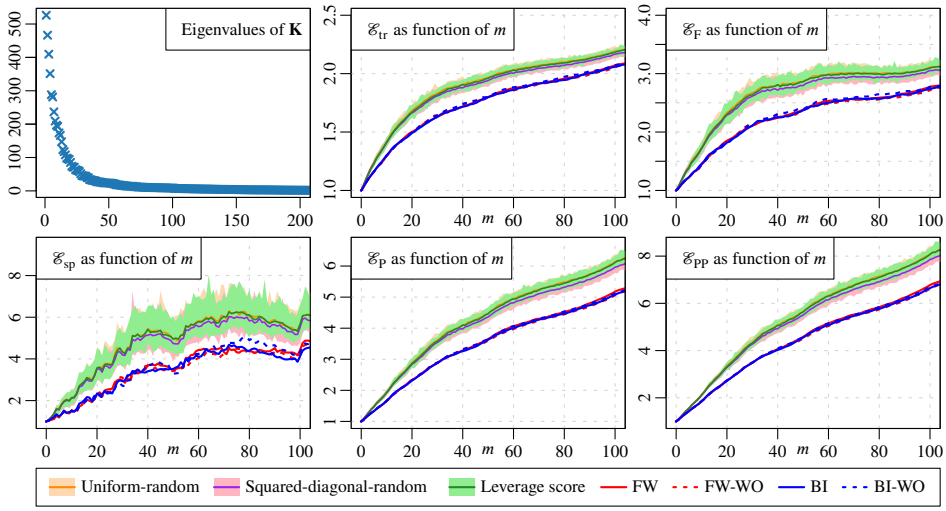

Figure 3. *For a random PSD matrix ($N = 1{,}500$), and for various sampling strategies, evolution of the five approximation factors (5.1) as functions of the number of columns m. The 200 largest eigenvalues of $\mathbf{K}$ are also displayed. For the stochastic methods, the solid line represents the median over 100 repetitions, and the boundaries of the shaded regions indicate the corresponding maximum and minimum values. The exact target potential $\mathfrak{g}$ is used. See Section 5.1 for more details.*

5.2.1. Exact target potential. We first consider the exact target potential $\mathfrak{g}$ and compare the accuracy of the Nyström approximations induced by four variants of Algorithm 1 (namely FW, BI, FW-WO, BI-WO) with the accuracy of the approximations obtained via uniform random sampling, leverage-score-based random sampling and k-DPP-based random sampling. The experiments involving random sampling are repeated 100 times. The results are presented in Figure 4, where we display the evolution of the approximation factors $\mathscr{E}_{F}$ and $\mathscr{E}_{P}$ up to $m = 100$ (the evolution of the other approximation factors is provided in Figure 7, in appendix; in terms of behaviour, $\mathscr{E}_{tr}$ and $\mathscr{E}_{sp}$ appear closely related to $\mathscr{E}_{F}$, while $\mathscr{E}_{PP}$ shows similarities with $\mathscr{E}_{P}$).

Remark 5.2. Following Remark 5.1, in Figure 4 (and in the complementary Figure 7, in appendix), to illustrate the decay of the spectrum of $\mathbf{K}$ we indicate the thresholds

$$\tau_{X} = \min\left\{m \in [N] \middle| \|\mathbf{K} - \mathbf{K}^{\star}_{m}\|_{X} \leqslant 0.01\|\mathbf{K}\|_{X}\right\}, X \in \{tr, F, sp\},$$

and with $\tau_{P} = \tau_{PP} = \tau_{F}$. For a given $X \in \{tr, F, sp, P, PP\}$, the smaller τ_{X} is, the faster the decay. ◁

In comparison to the considered random-sampling procedures, we observe that Algorithm 1 and its variants lead to more accurate approximations, especially in the range corresponding to the significant eigenvalues of $\mathbf{K}$ (this range is illustrated by the thresholds τ_{X} defined in Remark 5.2). After a certain number of iterations (which appears to be related to the spectrum of $\mathbf{K}$), the accuracy of the approximations induced by Algorithm 1 and its BI variant deteriorates (this is especially visible for $\gamma = 0.1$). The deterioration is stronger for $\mathscr{E}_{F}$ (and $\mathscr{E}_{tr}$ and $\mathscr{E}_{sp}$) than for $\mathscr{E}_{P}$ (and $\mathscr{E}_{PP}$), and the WO update rule appears to be able to mitigate this drop-off in accuracy (following Lemma 2.8, we recall that among the considered error maps, C_{P} and C_{PP} are the ones that are the most closely related to R). A comparison of the sample sizes required for random-uniform samples to achieve accuracies comparable to those of samples obtained via Algorithm 1 and its WO variant is provided in Figure 8 (in appendix).

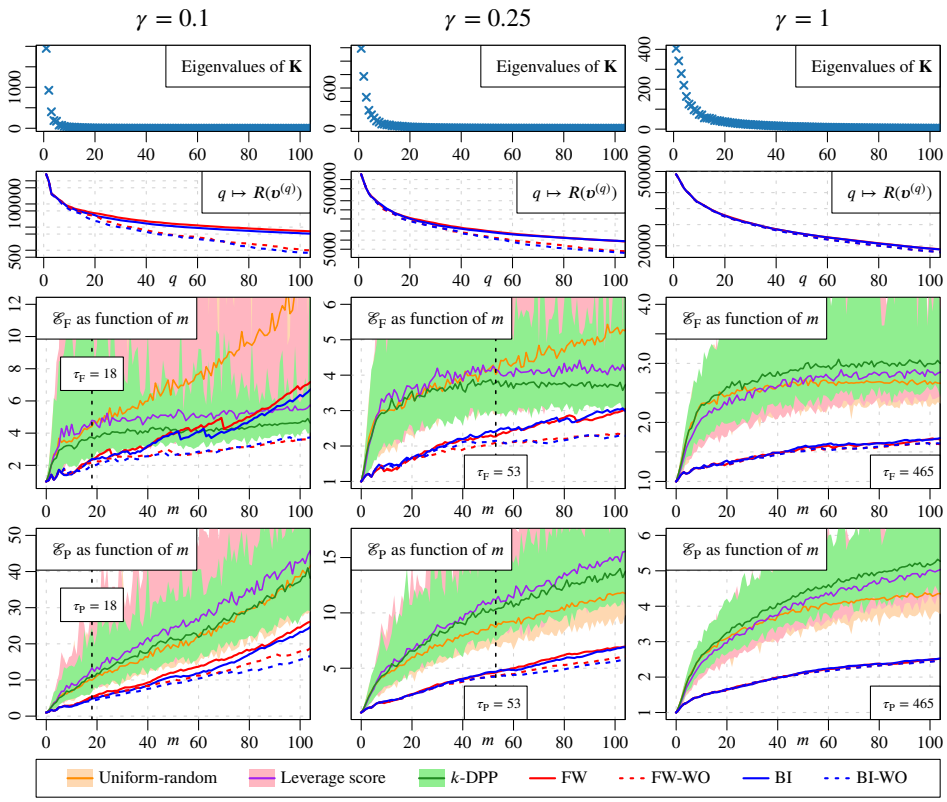

Figure 4. *For kernel matrices defined from the Abalone data set and squared exponential kernels, evolution of the approximation factors $\mathscr{E}_{\mathrm{F}}$ and $\mathscr{E}_{\mathrm{P}}$ as functions of the number of columns m (the evolution of the other approximation factors is provided in Figure 7, in appendix). 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 exact target potential* $\mathfrak{g}$ *is used). The evolutions 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 the stochastic strategies, we present the median, minimum and maximum of the approximation factors over* 100 *repetitions (see Figure 3). The vertical dashed lines indicate the value of the thresholds τ_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}\}$, defined in Remark 5.2 (when the threshold is outside the plot window, we only report its value). See Section 5.2.1 for more details.*

5.2.2. Approximate target potential. We now consider the stochastic variant S-MFW of Algorithm 1, that is, we use realisations of the estimator $\hat{\mathfrak{g}}_{\mathrm{sym}}$ (see Section 4) in combination with the modified FW direction (3.3), and we investigate the impact of the row-sample-size parameter ℓ on the accuracy of the induced Nyström approximations. For the kernel parameter, we use $\gamma = 0.25$ (intermediate case, see Figure 4) and we consider three different values of ℓ, namely $\ell = 100$, 250 and 500. The results are presented in Figure 5.

We observe that as ℓ increases, the accuracy of the Nyström approximations induced by the S-MFW procedure approaches that of the deterministic FW algorithm, and the variability in the approximation factors decreases. In the considered range of values of m, the obtained column samples maintain a high level of accuracy, even for small values of ℓ. Following Remarks 3.2 and 4.2, the maximum number of iterations of the S-MFW procedure tends to increase with ℓ. For this particular example,

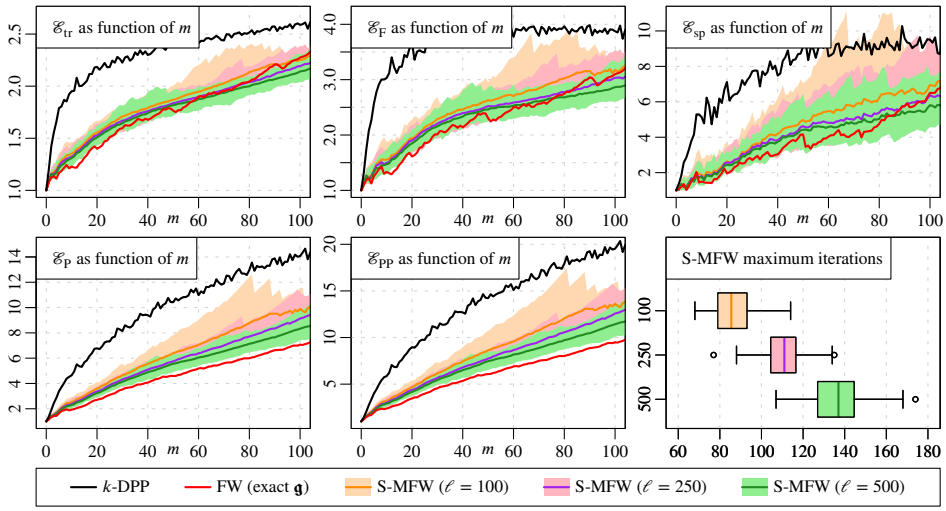

Figure 5. *For the kernel matrix defined from the Abalone data set and a squared exponential kernel with $\gamma = 0.25$, evolution of the five approximation factors $\mathscr{E}_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}, \mathrm{P}, \mathrm{PP}\}$, as functions of the number of columns m, for samples obtained using the S-MFW variant of Algorithm 1 (modified FW direction with realisations of $\hat{\mathfrak{g}}_{\mathrm{sym}}$; see Remark 3.2 and Section 4). Three different values of the row-sample-size parameter ℓ are considered. For each value of ℓ, we present the median, minimum and maximum of the approximation factors over 100 repetitions. For comparison, the approximation factors for the column samples obtained with Algorithm 1 (FW direction with exact target potential $\mathfrak{g}$) and through k-DPP-based random sampling (median over 100 repetitions) are also presented. The bottom-right plot displays the distribution of the maximum number of iterations of the S-MFW procedure for the considered values of ℓ (see Remark 3.2). The experiment is discussed in Section 5.2.2.*

considering $\ell = 500$ allows for a consistent exploration of the range $m \leqslant 100$ (see Section 5.3 for a further illustration of the link between ℓ and the maximum number of S-MFW iterations).

5.3. HIGGS data set. We now illustrate the ability of the proposed approach to handle large PSD matrices. We consider the HIGGS dataset (see [25]), 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 5.2) with $\gamma = 0.1$. To lower the memory requirement of the experiment, rather than being stored, the required entries of the matrix $\mathbf{K}$ are computed on demand from the data set and the kernel (*on-the-fly evaluation*).

In Figure 6, we display the decay of the error map R during the first 50,000 iterations of Algorithm 1 (exact target potential). Lemma 2.8 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 approximation $\hat{\mathbf{K}}(\boldsymbol{\upsilon}^{(q)})$ of $\mathbf{K}$ for $q = 1{,}000$; this approximation involves $m_q = 1{,}000$ columns of $\mathbf{K}$.

We next implement the S-MFW variant of Algorithm 1 for 10 realisations of the estimator $\hat{\mathfrak{g}}_{\mathrm{sym}}$ with $\ell = 10{,}000$. For these 10 realisations, the maximum number of S-MFW iterations is distributed between 65,000 and 67,000 (see Remark 3.2). We extract 10 samples of columns of size $m = 1{,}000$ and 2,000, and compare the trace errors of these samples with those of 10 random column samples of the same sizes (uniform sampling); the relatively small values of m are chosen to ensure a reasonably fast computation of the trace errors. The results are presented in Table 1.

As observed in Sections 5.1 and 5.2, the samples of columns obtained using Algorithm 1 and its S-MFW stochastic variant are noticeably more accurate than the ones obtained through random uniform

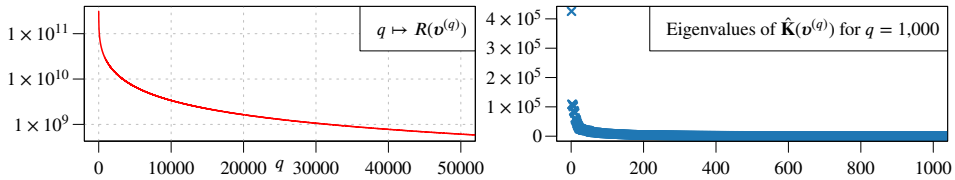

Figure 6. *For the HIGGS data set, decay of the the error map R during the* 50,000 *first iterations of Algorithm 1 (logarithmic scale). The non-zero eigenvalues of the Nyström approximation of* $\mathbf{K}$ *obtained at* $q = 1{,}000$ *are also presented. The experiment is discussed in Section 5.3.*

Table 1

For the HIGGS data set, summary statistics for the trace errors (rounded to the nearest integer) of various Nyström approximations of $\mathbf{K}$ *for* $m = 1{,}000$ *and* 2,000*. Results are presented for* 10 *random column samples (uniform sampling), and for* 10 *samples generated by the S-MFW variant of Algorithm 1 with* $\ell = 10{,}000$ *(stochastic approximations of* $\mathfrak{g}$*), as well as for the deterministic column samples produced by Algorithm 1 (exact target potential* $\mathfrak{g}$*). See Section 5.3 for more details.*

Method	**Number of columns** m	**Trace error**		
		Minimum	**Median**	**Maximum**
Uniform-random	1,000	7,090,945	7,117,127	7,149,980
	2,000	6,121,979	6,142,811	6,166,798
S-MFW ($\ell = 10{,}000$)	1,000	6,525,128	6,527,669	6,532,889
	2,000	5,698,986	5,703,138	5,707,372
FW (exact $\mathfrak{g}$)	1,000	—	6,439,653	—
	2,000	—	5,605,268	—

sampling, and for the considered values of m, the S-MFW variant is able to achieve an accuracy that is on par with the deterministic FW variant at a fraction of the numerical cost (here, $N/\ell = 1{,}100$).

6. Concluding discussion. We gave a detailed description of the framework surrounding the definition of a pseudoconvex differentiable relaxation of the CSP for PSD-matrix approximation, and described a class of gradient-based sequential sampling strategies leveraging the properties of this relaxation. The considered column-sampling procedures rely on the preliminary computation of a target potential, and stochastic approximation schemes can be implemented to reduce the time complexity of this operation. For PSD matrices of order N, and when relying on such stochastic approximations, the overall time complexity of the discussed strategies in linear in N. For instance, the worst-case time complexity of performing m iterations of the S-MFW variant of Algorithm 1 is $\mathcal{O}(m^2 + mN + \ell N)$, with in practice m and $\ell \ll N$; the algorithm then extracts a sample of m columns (and ℓ is the sample size used for the stochastic approximation of $\mathfrak{g}$).

We presented a series of experiments which demonstrate the ability of the proposed sampling strategies to produce accurate Nyström approximations while efficiently handling large PSD matrices. Notably, the discussed strategies appear to be able to achieve high levels of accuracy in ranges where other approaches (such as leverage-score and DPP-based sampling strategies) do not seem to lead to significant improvements over naive random column-sampling techniques, hence offering an interesting complement to the existing methodologies. The described procedures are in addition straightforward to implement, and the involved computations can be easily parallelised.

In view of our experiments, and especially for the optimal-step-size update rule, the range in which the discussed strategies are able to maintain high levels of accuracy appears to relate to the decay of the

spectrum of $\mathbf{K}$; gaining a deeper understanding of the mechanisms at play could improve the operating framework of the proposed procedures. In addition, although the error maps C_{X}, $\mathrm{X} \in \{\mathrm{sp}, \mathrm{F}, \mathrm{P}, \mathrm{PP}\}$, are upper-bounded by the surrogate error map R, obtaining tighter approximation bounds could help further support the considered relaxation. The impact of the stochastic approximation of the target potential on the column-sampling process could also warrant a more in-depth investigation. Finally, in complement to sequential sampling procedures, 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 techniques; see for instance [9, 11].

Acknowledgments. The authors thank the editors and the anonymous referee for their valuable comments and suggestions. M. Hutchings thankfully acknowledges funding from the Engineering and Physical Sciences Research Council grant EP/T517951/1.

Appendix A. Technical results.

In this section, we state and prove three technical lemmas. Following Remark 2.1, we introduce $\mathrm{HS}(\mathcal{E}, \mathcal{H}) = \{\mathbf{M} \in \mathbb{C}^{N\times N} \,|\, \mathrm{span}\{\mathbf{M}\} \subseteq \mathcal{H}\}$, and we set

$$\langle \mathbf{M} \,|\, \mathbf{T} \rangle_{\mathrm{HS}(\mathcal{E},\mathcal{H})} = \sum_{i\in[N]} \langle \mathbf{M}\mathbf{e}_i \,|\, \mathbf{T}\mathbf{e}_i \rangle_{\mathcal{H}} = \mathrm{trace}(\mathbf{M}^*\mathbf{K}^\dagger\mathbf{T}), \ \mathbf{M} \text{ and } \mathbf{T} \in \mathrm{HS}(\mathcal{E}, \mathcal{H}).$$

Endowed with the Hermitian form $\langle \cdot \,|\, \cdot \rangle_{\mathrm{HS}(\mathcal{E},\mathcal{H})}$, the linear space $\mathrm{HS}(\mathcal{E}, \mathcal{H})$ is a Hilbert space (indeed, we have $\|\mathbf{M}\|_{\mathrm{HS}(\mathcal{E},\mathcal{H})} = 0$ if and only $\mathbf{M}\mathbf{e}_i = 0$ for all $i \in [N]$, and so $\mathbf{M} = 0$).

Lemma A.1. *Let P and $Q \in \mathbb{C}^{N\times N}$ be two matrices corresponding to orthogonal projections onto closed linear subspaces of $\mathcal{H}$. We have $\|P\mathbf{K}Q\|^2_{\mathrm{HS}(\mathcal{H})} = \langle P\mathbf{K} \,|\, Q\mathbf{K} \rangle_{\mathrm{F}}$.*

Proof. We first observe that $P\mathbf{K} = \mathbf{K}P^* = P\mathbf{K}P^*$ (a similar property holds for Q). From (2.3), we indeed have

$$\begin{aligned}\mathbf{e}_i^* P\mathbf{K}\mathbf{e}_j &= \mathbf{e}_i^*\mathbf{K}\mathbf{K}^\dagger P\mathbf{K}\mathbf{e}_j = \langle \mathbf{K}\mathbf{e}_i \,|\, P\mathbf{K}\mathbf{e}_j \rangle_{\mathcal{H}} = \langle P\mathbf{K}\mathbf{e}_i \,|\, \mathbf{K}\mathbf{e}_j \rangle_{\mathcal{H}} = \langle P\mathbf{K}\mathbf{e}_i \,|\, \mathbf{e}_j \rangle_{\mathcal{E}} = \langle \mathbf{K}\mathbf{e}_i \,|\, P^*\mathbf{e}_j \rangle_{\mathcal{E}} \\ &= \langle \mathbf{K}\mathbf{e}_i \,|\, \mathbf{K}P^*\mathbf{e}_j \rangle_{\mathcal{H}} = \mathbf{e}_i^*\mathbf{K}\mathbf{K}^\dagger\mathbf{K}P^*\mathbf{e}_j = \mathbf{e}_i^*\mathbf{K}P^*\mathbf{e}_j, \ i \text{ and } j \in [N];\end{aligned}$$

in particular, the equality $\mathbf{e}_i^* P\mathbf{K}\mathbf{e}_j = \mathbf{e}_i^*\mathbf{K}\mathbf{K}^\dagger P\mathbf{K}\mathbf{e}_j$ follows by noticing that since $P\mathbf{K}\mathbf{e}_j \in \mathcal{H}$, there exists $\boldsymbol{\alpha} \in \mathbb{C}^N$ such that $P\mathbf{K}\mathbf{e}_j = \mathbf{K}\boldsymbol{\alpha}$. We then obtain

$$\begin{aligned}\|P\mathbf{K}Q\|^2_{\mathrm{HS}(\mathcal{H})} &= \mathrm{trace}(\mathbf{K}Q^*\mathbf{K}P^*\mathbf{K}^\dagger P\mathbf{K}Q) = \mathrm{trace}(\mathbf{K}Q^*P\mathbf{K}\mathbf{K}^\dagger\mathbf{K}P^*Q) \\ &= \mathrm{trace}(P\mathbf{K}P^*Q\mathbf{K}Q^*) = \mathrm{trace}(\mathbf{K}P^*Q\mathbf{K}),\end{aligned}$$

completing the proof. ∎

Lemma A.2. *For $\mathbb{J} \subseteq \mathbb{I} \subseteq [N]$, we have $\|\mathbf{K} - \hat{\mathbf{K}}(\mathbb{I})\|_{\mathrm{X}} \leqslant \|\mathbf{K} - \hat{\mathbf{K}}(\mathbb{J})\|_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$.*

Proof. Let $\mathcal{H}_{0\mathbb{I}}$ be the orthogonal complement of $\mathcal{H}_{\mathbb{I}}$ in $\mathcal{H}$; we set $P_{0\mathbb{I}} = \mathbf{I} - P_{\mathbb{I}}$. The matrix $P_{0\mathbb{I}}$ corresponds to the orthogonal projection from $\mathcal{H}$ onto $\mathcal{H}_{0\mathbb{I}}$ (and $\mathbf{K} - \hat{\mathbf{K}}(\mathbb{I}) = P_{0\mathbb{I}}\mathbf{K}$). We similarly introduce the subspace $\mathcal{H}_{0\mathbb{J}}$ and the matrix $P_{0\mathbb{J}}$. Since $\mathbb{J} \subseteq \mathbb{I}$, we have $\mathcal{H}_{\mathbb{J}} \subseteq \mathcal{H}_{\mathbb{I}}$, and we denote by $\mathcal{H}_e$ the orthogonal complement of $\mathcal{H}_{\mathbb{J}}$ in $\mathcal{H}_{\mathbb{I}}$; the matrix $P_e = P_{\mathbb{I}} - P_{\mathbb{J}}$ corresponds to the orthogonal projection from $\mathcal{H}$ onto $\mathcal{H}_e$.

Trace norm. Observing that $\langle P_e\mathbf{K} \mid P_{0\mathbb{I}}\mathbf{K}\rangle_{\mathrm{HS}(\mathcal{E},\mathcal{H})} = 0$, we have

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

Frobenius norm. Since $\mathcal{H}_{0\mathbb{I}}$ and $\mathcal{H}_e$ are orthogonal in $\mathcal{H}$, the matrices $P_{0\mathbb{I}}\mathbf{K}P_{0\mathbb{I}}$, $P_e\mathbf{K}P_e$, $P_{0\mathbb{I}}\mathbf{K}P_e$ and $P_e\mathbf{K}P_{0\mathbb{I}}$ are orthogonal in $\mathrm{HS}(\mathcal{H})$. Lemma A.1 then gives

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

Spectral norm. We first observe that if $P \in \mathbb{C}^{N\times N}$ is an orthogonal projection on $\mathcal{H}$, then the PSD operator on $\mathcal{E}$ related to $P\mathbf{K}$ and the PSD operator on $\mathcal{H}$ related to $P\mathbf{K}P$ have the same strictly-positive eigenvalues. Indeed, if $P\mathbf{K}\mathbf{v} = \lambda\mathbf{v}$, with $\mathbf{v} \in \mathcal{E}$, $\mathbf{v} \neq 0$ and $\lambda > 0$, then $PP\mathbf{K}\mathbf{v} = \lambda P\mathbf{v} = P\mathbf{K}\mathbf{v} = \lambda\mathbf{v}$, and so $\lambda(P\mathbf{v} - \mathbf{v}) = 0$; as $\lambda > 0$, we obtain $\mathbf{v} = P\mathbf{v} \in \mathcal{H}$ and $P\mathbf{K}P\mathbf{v} = \lambda\mathbf{v}$. Reciprocally, if $P\mathbf{K}P\boldsymbol{h} = \lambda\boldsymbol{h}$, with $\boldsymbol{h} \in \mathcal{H}$, $\boldsymbol{h} \neq 0$ and $\lambda > 0$, then $PP\mathbf{K}P\boldsymbol{h} = \lambda P\boldsymbol{h} = P\mathbf{K}P\boldsymbol{h} = \lambda\boldsymbol{h}$ and so $\lambda(P\boldsymbol{h} - \boldsymbol{h}) = 0$; as $\lambda > 0$, we have $P\boldsymbol{h} = \boldsymbol{h}$ and $P\mathbf{K}\boldsymbol{h} = \lambda\boldsymbol{h}$. Observing that $\mathcal{H}_{0\mathbb{I}} \subseteq \mathcal{H}_{0\mathbb{J}}$, we get

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

completing the proof. For the trace and Frobenius norms, an alternative characterisation of these inequalities can be found in [8, Lemma A.2]. ∎

Lemma A.3. *Let X and Y be two independent random variables following binomial distributions with size parameters m and $n \in \mathbb{N}$, respectively, and with same probability parameter $p \in [0,1]$. We have $\mathbb{E}(X|X+Y) = \frac{m}{m+n}(X+Y)$.*

Proof. We set $X = \sum_{i=1}^{m} B_i$ and $Y = \sum_{i=m+1}^{m+n} B_i$, with $\{B_i\}_{i\in[m+n]}$ a set of independent random variables following a Bernoulli distribution with parameter p. We have

$$X + Y = \mathbb{E}(X+Y|X+Y) = \sum_{i=1}^{m+n} \mathbb{E}(B_i|X+Y) = (m+n)\mathbb{E}(B_1|X+Y),$$

and $\mathbb{E}(X|X+Y) = \sum_{i=1}^{m} \mathbb{E}(B_i|X+Y) = m\mathbb{E}(B_1|X+Y)$. The result follows. ∎

Appendix B. Proofs.

This section gathers the proofs of the results presented in the main body of the paper.

Proof of Theorem 2.2. For $\boldsymbol{\xi} = \boldsymbol{\upsilon} + \rho(\boldsymbol{\eta} - \boldsymbol{\upsilon})$, $\boldsymbol{\upsilon}$ and $\boldsymbol{\eta} \in \mathbb{R}^N_{\geqslant 0}$, $\rho \in (0,1)$, we have $\mathbb{I}_{\boldsymbol{\xi}} = I_{\boldsymbol{\upsilon}} \cup I_{\boldsymbol{\eta}}$, and the maps $\rho \mapsto C_{\mathrm{X}}(\boldsymbol{\upsilon} + \rho[\boldsymbol{\eta} - \boldsymbol{\upsilon}])$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$, are thus constant on the open interval $(0,1)$. From Lemma A.2, we also have $C_{\mathrm{X}}(\boldsymbol{\xi}) \leqslant C_{\mathrm{X}}(\boldsymbol{\upsilon})$ and $C_{\mathrm{X}}(\boldsymbol{\xi}) \leqslant C_{\mathrm{X}}(\boldsymbol{\eta})$, concluding the proof. ∎

Proof of Theorem 2.5. We first show the quasiconvexity of R on $\mathbb{R}^N$. For $\boldsymbol{\xi} = \boldsymbol{\upsilon} + \rho(\boldsymbol{\eta} - \boldsymbol{\upsilon})$, $\boldsymbol{\upsilon}$ and $\boldsymbol{\eta} \in \mathbb{R}^N$, $\rho \in [0,1]$, there always 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 $\operatorname{coni}\{\boldsymbol{\upsilon}, \boldsymbol{\eta}\} = \operatorname{coni}\{c_{\boldsymbol{\upsilon}}\boldsymbol{\upsilon}, c_{\boldsymbol{\eta}}\boldsymbol{\eta}\}$ (with $\operatorname{coni}\{\boldsymbol{\upsilon}, \boldsymbol{\eta}\}$ the conical hull of $\{\boldsymbol{\upsilon}, \boldsymbol{\eta}\}$), so that $\boldsymbol{\xi} \in \operatorname{coni}\{c_{\boldsymbol{\upsilon}}\boldsymbol{\upsilon}, c_{\boldsymbol{\eta}}\boldsymbol{\eta}\}$ (in this case, $\boldsymbol{\xi} \in \mathscr{D}$ and $c > 0$).

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

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

and R is therefore quasiconvex on $\mathbb{R}^N$.

We now show the pseudoconvexity of R on $\mathscr{D}$. Let $\boldsymbol{\upsilon}$ and $\boldsymbol{\eta} \in \mathscr{D}$ be such that $\Theta(\boldsymbol{\upsilon}; \boldsymbol{\eta}) \geqslant 0$. As $\boldsymbol{\upsilon}^* \mathbf{S}(c_{\boldsymbol{\upsilon}}\boldsymbol{\upsilon} - \mathbb{1}) = 0$, the condition $\Theta(\boldsymbol{\upsilon}; \boldsymbol{\eta}) \geqslant 0$ reads $\boldsymbol{\eta}^* \mathbf{S}(c_{\boldsymbol{\upsilon}}\boldsymbol{\upsilon} - \mathbb{1}) \geqslant 0$, that is,

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

As $\boldsymbol{\upsilon}$ and $\boldsymbol{\eta} \in \mathscr{D}$, we have $\boldsymbol{\upsilon}^* \mathbf{S}\mathbb{1} > 0$, $\boldsymbol{\upsilon}^* \mathbf{S}\boldsymbol{\upsilon} > 0$ and $\boldsymbol{\eta}^* \mathbf{S}\mathbb{1} > 0$, and so, from (B.1), $\boldsymbol{\eta}^* \mathbf{S}\boldsymbol{\upsilon} > 0$. The matrix $\mathbf{S}$ being PSD, the CS 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 (B.1), we get (note that we also have $\boldsymbol{\eta}^* \mathbf{S}\boldsymbol{\eta} > 0$ as $\boldsymbol{\eta} \in \mathscr{D}$)

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

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

Proof of Lemma 2.6. We set $a = \boldsymbol{\mathfrak{g}}^* \boldsymbol{\upsilon} > 0$, $b = \boldsymbol{\mathfrak{g}}^* \boldsymbol{\eta}$, $c = \boldsymbol{\upsilon}^* \mathbf{S}\boldsymbol{\upsilon} > 0$, $d = \boldsymbol{\eta}^* \mathbf{S}\boldsymbol{\eta}$ and $e = \boldsymbol{\upsilon}^* \mathbf{S}\boldsymbol{\eta}$. For $x \in \mathbb{R}$, we also set $\boldsymbol{\xi}_x = \boldsymbol{\upsilon} + x(\boldsymbol{\eta} - \boldsymbol{\upsilon})$, and we introduce the functions

$$\varphi(x) = \boldsymbol{\mathfrak{g}}^* \boldsymbol{\xi}_x = x(b-a) + a \quad \text{and} \quad \psi(x) = \boldsymbol{\xi}_x^* \mathbf{S}\boldsymbol{\xi}_x = x^2(c+d-2e) + 2x(e-c) + c.$$

The condition $\Theta(\boldsymbol{\upsilon}; \boldsymbol{\eta}) < 0$ ensures that the degree-2 polynomial ψ is strictly positive; indeed, ψ is non-negative and admits a real root if and only if $e^2 = cd$, that is, from the CS inequality, if $\boldsymbol{\eta} = \alpha\boldsymbol{\upsilon} + \boldsymbol{\epsilon}$, with $\alpha \in \mathbb{R}$ and $\boldsymbol{\epsilon} \in \mathbb{R}^N$ such that $\mathbf{S}\boldsymbol{\epsilon} = 0$, and we would in this case have $\Theta(\boldsymbol{\upsilon}; \boldsymbol{\eta}) = 0$.

We define $f(x) = -\varphi^2(x)/\psi(x)$, $x \in \mathbb{R}$; if $\boldsymbol{\xi}_x \in \mathscr{D}$, then $f(x) = R(\boldsymbol{\xi}_x) - \|\mathbf{K}\|_{\mathrm{F}}^2$. We have

$$f'(x) = 2\frac{\varphi(x)}{\psi^2(x)}\Big[x\big((bc-ae) + (ad-be)\big) - (bc-ae)\Big], x \in \mathbb{R},$$

so that f admits at most two stationary points on $\mathbb{R}$. The conditions on $\boldsymbol{\upsilon}$ and $\boldsymbol{\eta}$ and the pseudoconvexity of R on $\mathscr{D}$ ensure that the function $\rho \mapsto R(\boldsymbol{\xi}_\rho)$ admits a minimum on $(0,1]$; the argument of this minimum is the optimal step size r and corresponds to a stationary point of f. If $a = b$, the function φ is constant and strictly positive (as $a > 0$). If $a \neq b$, for $x_1 = a/(a-b)$, we have $\varphi(x_1) = 0$, and so $f'(x_1) = 0$. However, we then have $\boldsymbol{\mathfrak{g}}^* \boldsymbol{\xi}_{x_1} = 0$, and so $R(\boldsymbol{\xi}_{x_1}) = \|\mathbf{K}\|_{\mathrm{F}}^2 > R(\boldsymbol{\upsilon})$; we can therefore

conclude that $r \neq x_1$. Canceling the linear function $x \mapsto x\big((bc - ae) + (ad - be)\big) - (bc - ae)$, we obtain $f'(x_2) = 0$ with

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

and so $r = x_2$; we then have $\mathcal{I}(\upsilon; \boldsymbol{\eta}) = f(0) - f(x_2) = (bc - ae)^2 / \big(c(cd - e^2)\big)$. ∎

Proof of Proposition 2.7. We follow the proof of Theorem 2.2, and show that if $\mathbb{J} \subseteq \mathbb{I} \subseteq [N]$, then

$$\|\mathbf{K} - P_{\mathbb{I}}\mathbf{K}\|_{\mathrm{HS}(\mathcal{H})} \leqslant \|\mathbf{K} - P_{\mathbb{J}}\mathbf{K}\|_{\mathrm{HS}(\mathcal{H})} \quad \text{and} \quad \|\mathbf{K} - P_{\mathbb{I}}\mathbf{K}P_{\mathbb{I}}\|_{\mathrm{HS}(\mathcal{H})} \leqslant \|\mathbf{K} - P_{\mathbb{J}}\mathbf{K}P_{\mathbb{J}}\|_{\mathrm{HS}(\mathcal{H})};$$

see [8, Section 4] for an alternative characterisation of these inequalities. Using the same notations as in the proof of Lemma A.2 and noticing that $\mathcal{H}_{0\mathbb{I}}$ and $\mathcal{H}_e$ are orthogonal in $\mathcal{H}$, we have

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

as required. Next, if P is an orthogonal projection on $\mathcal{H}$, then $\langle \mathbf{K} \mid P\mathbf{K}P\rangle_{\mathrm{HS}(\mathcal{H})} = \|P\mathbf{K}P\|^2_{\mathrm{HS}(\mathcal{H})}$ and

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

Observing that the matrices $P_{\mathbb{J}}\mathbf{K}P_{\mathbb{J}}$, $P_e\mathbf{K}P_e$, $P_{\mathbb{J}}\mathbf{K}P_e$ and $P_e\mathbf{K}P_{\mathbb{J}}$ are orthogonal in $\mathrm{HS}(\mathcal{H})$, we obtain

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

giving, in combination with (B.2), the expected inequality. ∎

Proof of Lemma 2.8. The inequality $C_{\mathrm{sp}}(\upsilon) \leqslant C_{\mathrm{F}}(\upsilon)$ follows from the relation between the Frobenius and spectral norms. From Lemma A.1, we have (with $\Re(z)$ the real part of $z \in \mathbb{C}$)

$$C_{\mathrm{F}}(\upsilon) = \|\mathbf{K}\|^2_{\mathrm{F}} + \|\hat{\mathbf{K}}(\upsilon)\|^2_{\mathrm{F}} - 2\Re\big(\langle \mathbf{K} \mid \hat{\mathbf{K}}(\upsilon)\rangle_{\mathrm{F}}\big) = \|\mathbf{K}\|^2_{\mathrm{F}} + \|P_\upsilon\mathbf{K}P_\upsilon\|^2_{\mathrm{HS}(\mathcal{H})} - 2\|P_\upsilon\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{H})}. \tag{B.3}$$

We introduce $P_{0\upsilon} = \mathbf{I} - P_\upsilon$. The matrix $P_{0\upsilon}$ correspond to the orthogonal projection from $\mathcal{H}$ onto the orthogonal complement of $\mathcal{H}_\upsilon$ in $\mathcal{H}$, and so

$$\|P_\upsilon\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{H})} = \|P_\upsilon\mathbf{K}P_\upsilon\|^2_{\mathrm{HS}(\mathcal{H})} + \|P_\upsilon\mathbf{K}P_{0\upsilon}\|^2_{\mathrm{HS}(\mathcal{H})} \geqslant \|P_\upsilon\mathbf{K}P_\upsilon\|^2_{\mathrm{HS}(\mathcal{H})}. \tag{B.4}$$

Combining (B.3) and (B.4), we obtain

$$C_{\mathrm{F}}(\upsilon) \leqslant \|\mathbf{K}\|^2_{\mathrm{F}} - \|P_\upsilon\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{H})} = C_{\mathrm{P}}(\upsilon) \leqslant \|\mathbf{K}\|^2_{\mathrm{F}} - \|P_\upsilon\mathbf{K}P_\upsilon\|^2_{\mathrm{HS}(\mathcal{H})} = C_{\mathrm{PP}}(\upsilon).$$

We next observe that $\mathbf{K}\mathbf{V}\boldsymbol{h} = P_\upsilon\mathbf{K}\mathbf{V}P_\upsilon\boldsymbol{h}$, $\boldsymbol{h} \in \mathcal{H}$ (indeed, we have $\mathrm{span}\{\mathbf{K}\mathbf{V}\} \subseteq \mathcal{H}_\upsilon$, and $\mathbf{e}_i^* P_\upsilon \boldsymbol{h} = \mathbf{e}_i^* \boldsymbol{h}$ for all $i \in \mathbb{I}_\upsilon$), and so $\langle \mathbf{K} - P_\upsilon\mathbf{K}P_\upsilon \mid P_\upsilon\mathbf{K}P_\upsilon - \mathbf{K}\mathbf{V}\rangle_{\mathrm{HS}(\mathcal{H})} = 0$. We hence obtain

$$\|\mathbf{K} - \mathbf{K}\mathbf{V}\|^2_{\mathrm{HS}(\mathcal{H})} = \|\mathbf{K} - P_\upsilon\mathbf{K}P_\upsilon\|^2_{\mathrm{HS}(\mathcal{H})} + \|P_\upsilon\mathbf{K}P_\upsilon - \mathbf{K}\mathbf{V}\|^2_{\mathrm{HS}(\mathcal{H})},$$

and so $C_{\mathrm{PP}}(\upsilon) \leqslant D(\upsilon)$. Observing that $C_{\mathrm{PP}}(\upsilon) \leqslant \|\mathbf{K}\|^2_{\mathrm{F}} = R(0)$ and that $R(\upsilon) = \min_{c \geqslant 0} D(c\upsilon)$, we necessarily have $C_{\mathrm{PP}}(\upsilon) \leqslant R(\upsilon) \leqslant D(\upsilon)$, completing the expected sequence of inequalities.

We conclude the proof by observing that if $\mathbf{S}_{i,i} > 0$, $i \in [N]$, then $\|\hat{\mathbf{K}}(\mathbf{e}_i)\|^2_{\mathrm{F}} = (\mathfrak{g}^*\mathbf{e}_i)^2/\mathbf{S}_{i,i}$, and if $\mathbf{S}_{i,i} = 0$, then $\|\hat{\mathbf{K}}(\mathbf{e}_i)\|^2_{\mathrm{F}} = 0$ and $\mathbf{e}_i \notin \mathscr{D}$. ∎

Appendix C. Abalone data set: additional figures.

In this section, we further illustrate the results of our experiments on kernel matrices defined from the Abalone data set (Section 5.2).

C.1. Complement to Figure 4. Figure 7 complements Figure 4 by providing the evolution, as functions of the number of columns m, of the approximation factors $\mathscr{E}_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{sp}, \mathrm{PP}\}$, for the various sampling strategies considered in Section 5.2.1 (exact target potential $\mathfrak{g}$).

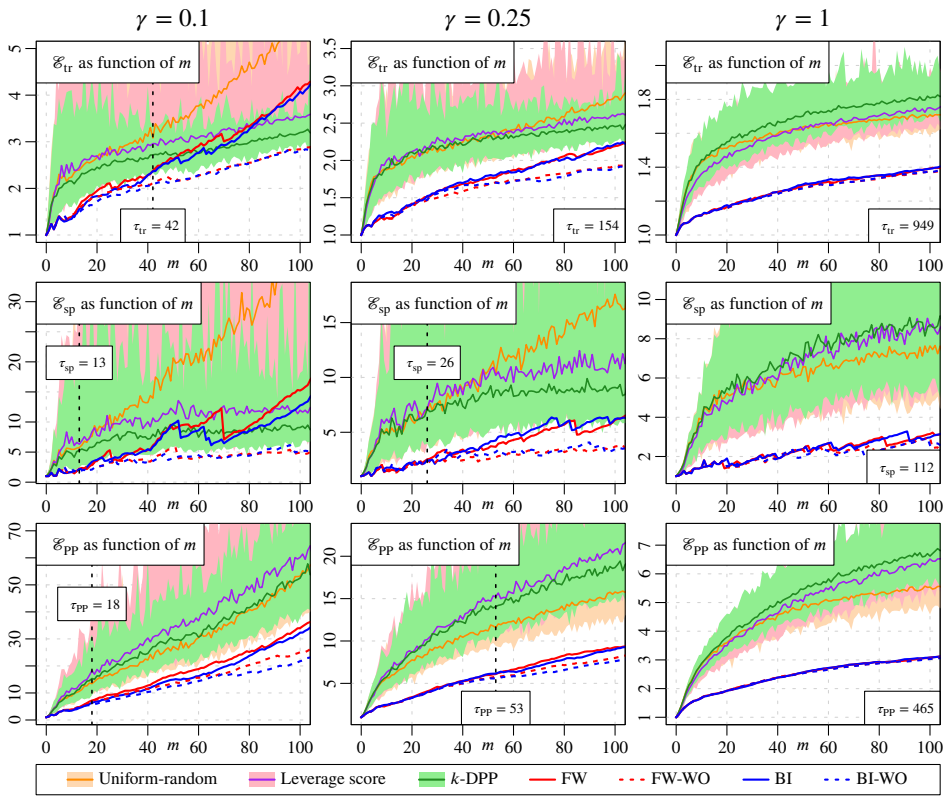

Figure 7. *In complement to Figure 4 and for the various sampling strategies considered in Section 5.2.1, evolution of the approximation factors $\mathscr{E}_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{sp}, \mathrm{PP}\}$, as functions of the number of columns m (Abalone data set and squared-exponential kernel). The values of the corresponding thresholds τ_{X}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{sp}, \mathrm{PP}\}$ are also indicated (see Remark 5.2).*

C.2. Approximation accuracy versus sample sparsity. In Figure 8, and in the framework of Figure 4 (Abalone data set and exact target potential $\mathfrak{g}$, see Section 5.2.1), we compare the sample sizes required for random-uniform samples to achieve accuracies comparable to those of samples obtained via Algorithm 1 and its WO variant. For simplicity, we solely consider the error map C_{F} (a similar behaviour is nevertheless also observed for the error maps C_{X}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{sp}, \mathrm{P}, \mathrm{PP}\}$).

For $\alpha \in [0, C_{\mathrm{F}}(0)]$, we denote by $m_{\mathrm{FW}}(\alpha)$ the minimum sample size required for a sample $\boldsymbol{I}_{\boldsymbol{\upsilon}} \subseteq [N]$ obtained with Algorithm 1 to achieve $C_{\mathrm{F}}(\boldsymbol{\upsilon}) \leqslant \alpha$. We similarly define the sample sizes $m_{\text{FW-WO}}(\alpha)$ for the WO variants of Algorithm 1, and $m_{\mathrm{unif}}(\alpha)$ for random-uniform sampling (in this case, the median of $C_{\mathrm{F}}(\boldsymbol{\upsilon})$ over 100 repetitions is considered). A schematic illustration of the definition of $m_{\mathrm{FW}}(\alpha)$, $m_{\text{FW-WO}}(\alpha)$ and $m_{\mathrm{unif}}(\alpha)$ is provided in Figure 8. We then represent the evolution, as a function of m, of the sample-size ratio $m_{\mathrm{unif}}(\alpha)/m_{\mathrm{FW}}(\alpha)$, with α such that $m_{\mathrm{FW}}(\alpha) = m$. The evolution of the sample-size ratio $m_{\mathrm{unif}}(\alpha)/m_{\text{FW-WO}}(\alpha)$ is presented accordingly. In the considered range of values of $m_{\mathrm{FW}}(\alpha)$ and $m_{\text{FW-WO}}(\alpha)$ (that is, between 1 and 100), the observed sample-size ratios vary between 1.47 and 3.8.

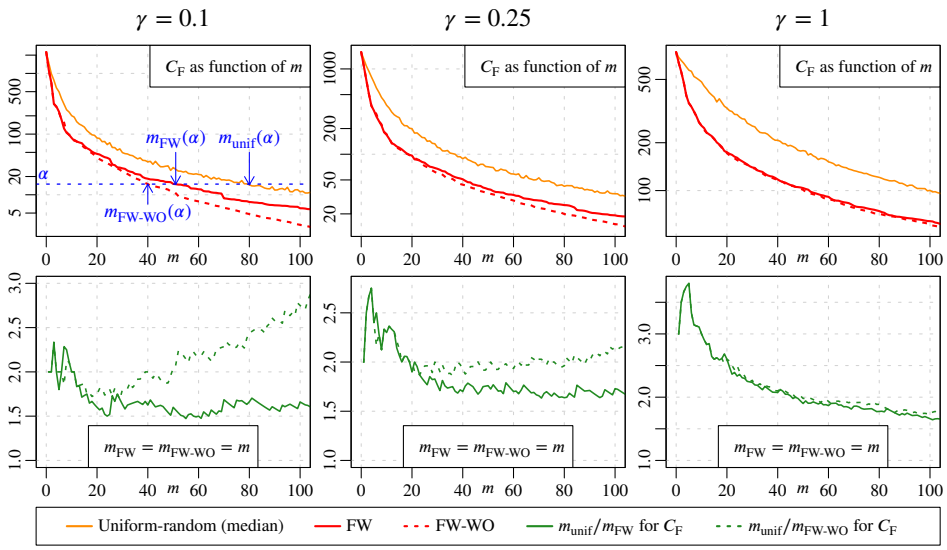

Figure 8. *In the framework of Section 5.2.1 (Abalone data set and exact target potential) and in complement to Figure 4, evolution of the error map C_F (logarithmic scale) as a function of the number of columns m, for samples obtained using random-uniform sampling (median over* 100 *repetitions), and for Algorithm 1 and its WO variant (top). Also, comparison of the sample sizes required for random-uniform sampling to achieve accuracies similar to those of the samples obtained via Algorithm 1 and its WO variant (bottom; for simplicity, references to α are omitted in the notations); see Section C.2. Each column in the figure corresponds to a different value of the kernel parameter γ.*

REFERENCES

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

[2] F. Bach, S. Lacoste-Julien, and G. Obozinski, *On the equivalence between herding and conditional gradient algorithms*, in International Conference on Machine Learning, 2012.

[3] M.-A. Belabbas and P. J. Wolfe, *Spectral methods in machine learning and new strategies for very large datasets*, Proc. Natl. Acad. Sci. USA, 106 (2009), pp. 369–374.

[4] Y. Chen, M. Welling, and A. J. Smola, *Super-samples from kernel herding*, in Uncertainty in Artificial Intelligence, 2010.

[5] M. Derezinski, R. Khanna, and M. W. Mahoney, *Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nyström method*, in Advances in Neural Information Processing Systems, vol. 33, 2020.

[6] M. Derezinski and M. W. Mahoney, *Determinantal point processes in randomized numerical linear algebra*, Notices Amer. Math. Soc., 68 (2021), pp. 34–45.

[7] P. Drineas and M. W. Mahoney, *On the Nyström method for approximating a Gram matrix for improved kernel-based learning*, J. Mach. Learn. Res., 6 (2005), pp. 2153–2175.

[8] B. Gauthier, *Kernel embedding of measures and low-rank approximation of integral operators*, Positivity, 28 (2024).

[9] B. Gauthier and J. Suykens, *Optimal quadrature-sparsification for integral operator approximation*, SIAM J. Sci. Comput., 40 (2018), pp. A3636–A3674.

[10] A. Gittens and M. W. Mahoney, *Revisiting the Nyström method for improved large-scale machine learning*, J. Mach. Learn. Res., 17 (2016), pp. 1–65.

[11] M. Hutchings and B. Gauthier, *Local optimisation of Nyström samples through stochastic gradient descent*, in Machine Learning, Optimization, and Data Science, Springer, 2023, pp. 123–140.

[12] S. Kumar, M. Mohri, and A. Talwalkar, *Sampling methods for the Nyström method*, J. Mach. Learn. Res., 13 (2012), pp. 981–1006.

[13] S. Lacoste-Julien, F. Lindsten, and F. Bach, *Sequential kernel herding: Frank-wolfe optimization for particle*

filtering, in Artificial Intelligence and Statistics, PMLR, 2015.
[14] C. Li, S. Jegelka, and S. Sra, *Fast DPP sampling for Nyström with application to kernel methods*, in International Conference on Machine Learning, vol. 48, PMLR, 2016, pp. 2061–2070.
[15] F. Mezzadri, *How to generate random matrices from the classical compact groups*, Notices Amer. Math. Soc., 54 (2007), pp. 592–604.
[16] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, *Kernel mean embedding of distributions: a review and beyond*, Found. Trends Mach. Learn., 10 (2017), pp. 1–141.
[17] C. Musco and C. Musco, *Recursive sampling for the Nyström method*, in Advances in Neural Information Processing Systems, vol. 30, 2017.
[18] W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford, *Abalone*. UCI Machine Learning Repository, 1995.
[19] V. I. Paulsen and M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge University Press, 2016.
[20] E. Pauwels, F. Bach, and J.-P. Vert, *Relating leverage scores and density using regularized Christoffel functions*, in Advances in Neural Information Processing Systems, vol. 31, 2018.
[21] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT press, Cambridge, MA, 2006.
[22] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet, *On the empirical estimation of integral probability metrics*, Electron. J. Stat., 6 (2012), pp. 1550–1599.
[23] N. Sterge, B. Sriperumbudur, L. Rosasco, and A. Rudi, *Gain with no pain: efficiency of kernel-PCA by Nyström sampling*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 3642–3652.
[24] S. Wang, A. Gittens, and M. W. Mahoney, *Scalable kernel k-means clustering with Nyström approximation: relative-error bounds*, J. Mach. Learn. Res., 20 (2019), pp. 431–479.
[25] D. Whiteson, *HIGGS*. UCI Machine Learning Repository, 2014.
[26] C. Williams and M. Seeger, *Using the Nyström method to speed up kernel machines*, in Advances in Neural Information Processing Systems, vol. 13, 2000, pp. 682–688.