Energy-based sampling strategies for the low-rank approximation of positive semidefinite matrices

Matthew James Hutchings
School of Mathematics
Cardiff University

A thesis submitted for the degree of
Doctor of Philosophy

June 2024

Abstract

This thesis focuses on the design of efficient strategies for the low-rank approximation of positive semidefinite matrices via column sampling. A special emphasis is placed on investigating the properties of the energy setting, which relates the low-rank approximation of Hilbert-Schmidt integral operators with the approximation of potentials in reproducing kernel Hilbert spaces. The implications of the energy setting in the matrix framework are investigated, leading to the definition of differentiable surrogate error maps for the characterisation of low-rank approximations. Classes of gradient-based sampling strategies leveraging the properties of these error maps are then proposed and analysed, and the possibility to improve the numerical efficiency of these approaches via stochastic approximations is explored.

Statement of originality

I declare that this thesis is the result of my own work, except where otherwise indicated in the text. The energy setting described in Chapter 3 is based on the work of Gauthier (2024) by my doctoral supervisor Dr. Bertrand Gauthier (BG). The theory and methodology for the sequential sampling strategies described in Chapter 4 were developed in collaboration with BG, and the experiments of Chapter 5 were conducted jointly with BG.

Chapter 4 of the thesis is adapted from the work of Hutchings and Gauthier (2023a), entitled *Energy-based sequential sampling for low-rank PSD-matrix approximation*, to appear in the SIAM Journal of Optimisation and Data Science (SIMODS). Chapter 5 is adapted from the work of Hutchings and Gauthier (2023b), entitled *Local optimisation of Nyström samples through stochastic gradient descent*, published in the LOD 2022 Conference proceedings.

Acknowledgments

Firstly, I would like to thank my academic supervisors Dr. Bertrand Gauthier and Dr. Kirstin Strokorb for their guidance and support; the skills you have taught me have been invaluable, and I have grown so much as a researcher thanks to you. I also thank the Engineering and Physical Sciences Research Council for their financial support which has made this research possible.

I am grateful to the School of Mathematics at Cardiff University and our wonderful PGR community for fostering a stimulating and encouraging research environment. To my fellow Cardiff SIAM-IMA Chapter members and organisers, thank you for your help and participation in all of our social and academic events over the years. To my friends Michela, Matthew and Elizabeth, thank you for the countless hours of support and entertainment, making my time in the office so enjoyable.

Finally, thank you Mum and Dad for your endless love and encouragement which has motivated me throughout this journey. I could not have done this without you.

Dedicated to Peter Marley Keen,

my inspiration.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Main contributions and organisation of the manuscript 2

2 Low-rank approximation of positive semidefinite matrices **4**

2.1 Low-rank approximation: general case 4

2.1.1 Approximation accuracy 4

2.1.2 Rank-optimal approximations 6

2.1.3 Low-rank approximation via column sampling 7

2.2 Positive-semidefinite matrices 8

2.2.1 Nyström approximation 8

2.2.2 The column-sampling problem 9

2.2.3 Approaches to the column-sampling problem 10

2.2.3.1 Naive sampling strategies 10

2.2.3.2 Leverage scores 11

2.2.3.3 Determinantal point processes 13

3 Reproducing kernel Hilbert spaces and kernel energies **15**

3.1 Hilbert–Schmidt operators and reproducing kernel Hilbert spaces . 15

3.1.1 Reproducing kernel Hilbert spaces 15

3.1.2 Hilbert–Schmidt operators 16

3.1.3 Integral operators and the energy setting 17

3.1.3.1 Integral operators with PSD kernels 18

3.1.3.2 Potentials in squared-kernel RKHSs 19

3.2 Reproducing kernel Hilbert spaces and PSD matrix approximation . 20

3.2.1 Nyström approximations and projections 21

3.2.2 Error maps and Hilbert–Schmidt norms 24

4 Sequential column sampling strategies 25

4.1 Relaxing the column-sampling problem 25

4.1.1 First relaxation: selection vectors 26

4.1.2 Second relaxation: quadrature approximation 28

4.1.3 Invariance under rescaling 29

4.1.4 Additional error maps and further properties 33

4.2 Gradient-based sequential sampling 34

4.3 Stochastic approximation of the target potential 38

4.4 Numerical experiments 41

4.4.1 Random PSD matrix 42

4.4.2 Abalone data set 43

4.4.2.1 Exact target potential 44

4.4.2.2 Approximate target potential 45

4.4.3 HIGGS data set 47

Appendix: Additional figure for Abalone data set 49

5 Particle flow-based sampling strategies 50

5.1 Approximation of PSD kernel matrices 50

5.1.1 Nyström approximation through landmark points 51

5.1.2 Approximation accuracy 52

5.2 An energy-based error map 53

5.3 A convergence result 55

5.4 Stochastic approximation of the gradient 56

5.5 Numerical experiments 58

5.5.1 Bi-Gaussian example 59

5.5.2 Abalone data set 62

5.5.3 MAGIC data set 63

5.5.4 MiniBooNE data set 64

Appendix: Proof of Theorem 5.1 65

6 Concluding discussion **68**

6.1 Summary of the contributions of the thesis 68

6.2 Extensions and future work 69

6.2.1 Extensions of the presented work 70

6.2.2 Other research directions 70

Bibliography **72**

List of Figures

4.1 Schematic representation of the error maps D, R and C_{F} on $\mathbb{R}^N_{\geqslant 0}$ for an example 2×2 PSD matrix $\mathbf{K}$. 30

4.2 For the random PSD matrix example of Section 4.4.1, evolution of the value of the error maps R and C_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}, \mathrm{PP}\}$, during the first 100 iterations of Algorithm 4.1 (left) and its BI variant (right). 42

4.3 For the random PSD matrix example of Section 4.4.1, and for various sampling strategies, evolution of the five approximation factors (4.13) as functions of the number of sampled columns. The 200 largest eigenvalues of $\mathbf{K}$ are also displayed. 43

4.4 For the Abalone data set example of Section 4.4.2, evolution of the approximation factors $\mathscr{E}_{\mathrm{F}}$ and $\mathscr{E}_{\mathrm{P}}$ as functions of the number of sampled columns, for different values of the kernel parameter γ and for various sampling strategies. 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 first 100 iterations of the four considered variants of Algorithm 4.1. The values of the thresholds τ_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}\}$, are also indicated. 45

4.5 For the Abalone data set example of Section 4.4.2 with kernel parameter $\gamma = 0.25$, evolution of the five approximation factors (4.13) as functions of the number of sampled columns, for samples obtained using Algorithm 4.1 and its S-MFW variant, as well as for samples obtained through k-DPP-based random sampling. For the S-MFW variant, three different values of the row-sample-size parameter ℓ are considered, and the bottom-right plot displays the distribution of the maximum number of iterations of the S-MFW procedure for each value of ℓ. 46

4.6 For the HIGGS data set example of Section 4.4.3, decay of the the error map R during the first 50,000 iterations of Algorithm 4.1 (logarithmic scale). The non-zero eigenvalues of the Nyström approximation of $\mathbf{K}$ obtained at $q = 1{,}000$ are also presented. 47

4.7 For the Abalone data set example of Section 4.4.2, and in complement to Figure 4.4, evolution of the approximation factors $\mathscr{E}_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{sp}, \mathrm{PP}\}$, as functions of the number of sampled columns for the various sampling strategies considered in Section 4.4.2.1; the values of the corresponding thresholds τ_{X}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{sp}, \mathrm{PP}\}$, are also indicated. 49

5.1 For the bi-Gaussian example of Section 5.5.1, graphical representation of the path $t \mapsto \mathcal{S}^{(t)}$ followed by the landmark points of a Nystrom sample during the local minimisation of $\mathfrak{R}$ through GD (left). The evolution, during the GD, of $\mathfrak{R}$ and the trace error map is also presented (right). 60

5.2 For the bi-Gaussian example of Section 5.5.1, and for different Nyström sample sizes, comparison of the values of $\mathfrak{R}$ and of the approximation factors (5.7) for the initial random samples $\mathcal{S}^{(0)}$ and the locally optimised samples $\mathcal{S}^{(T)}$ obtained through GD. 61

5.3 For the bi-Gaussian example of Section 5.5.1, graphical representation of the paths followed by the landmark points of a random initial Nyström sample during the local minimisations of $\mathfrak{R}$ and $\mathfrak{C}_{\mathrm{tr}}$ through GD (left). For a set of random initial Nyström samples, comparison of the improvements yielded by the minimisations of $\mathfrak{R}$ and $\mathfrak{C}_{\mathrm{tr}}$ in terms of $\mathfrak{R}$ (middle) and trace norm (right). 62

5.4 For the Abalone data set example of Section 5.5.2, and for different values of the kernel parameter, comparison of the values of $\mathfrak{R}$ and of the approximation factors (5.7) for a set of initial Nyström samples $\mathcal{S}^{(0)}$ and the locally optimised samples $\mathcal{S}^{(T)}$ obtained through SGD with i.i.d. sampling. 63

5.5 For the MAGIC data set example of Section 5.5.3, and for different Nyström sample sizes, boxplots of the error map $\mathfrak{R}$ and of the approximation factor $\mathfrak{E}_{\mathrm{tr}}$ before and after the local optimisation via SGD of a set of random Nyström samples (left). A graphical representation of the decay of $\mathfrak{R}$ is also presented (right). 64

5.6 For the MiniBooNE data set of Section 5.5.4, decay of the error map $\mathfrak{R}$ during the optimisation of a random initial Nyström sample. . . 65

List of Tables

2.1 Computational complexities of evaluating the trace, Frobenius and spectral norms of the approximation error for a Nyström approximation $\hat{\mathbf{K}}(I)$ with $|I| = m \leqslant N$. 10

4.1 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 4.1 with $\ell = 10{,}000$ (stochastic approximations of $\mathfrak{g}$), as well as for the deterministic column samples produced by Algorithm 4.1 (exact target potential $\mathfrak{g}$). 48

Commonly-used notations

$\mathbb{N}$	The set of natural numbers, the strictly positive integers.
$[N]$	The set of positive integers between 1 and N, inclusive.
$\mathbb{R}$, $\mathbb{C}$	The sets of real and complex numbers, respectively.
$\mathbb{R}_{\geqslant 0}$	The set of non-negative real numbers.
$\mathbb{F}^N$	The set of vectors of size N which have elements in a set $\mathbb{F}$.
$\mathbb{F}^{N_\mathrm{r} \times N_\mathrm{c}}$	The set of matrices with N_r rows and N_c columns which have elements in a set $\mathbb{F}$.
$\mathbb{1}$	The vector with all elements equal to 1.
$\lvert S \rvert$	The cardinality of a set S.
2^S	The power set of a set S.
$\lvert z \rvert$	The modulus of a complex number z (absolute value for reals).
$\overline{z}$	The conjugate of a complex number z.
$\mathrm{Re}(z)$	The real part of a complex number z.
$\mathbf{I}$	The identity matrix.
$\mathbf{M}^T$	The transpose of a matrix $\mathbf{M}$.
$\overline{\mathbf{M}}$	The conjugate of a matrix $\mathbf{M}$ (elementwise conjugation).
$\mathbf{M}^*$	The conjugate transpose of a matrix $\mathbf{M}$.
$\mathbf{M}_{I,\bullet}$	The rows of a matrix $\mathbf{M}$ with indices in I.
$\mathbf{M}_{\bullet,J}$	The columns of a matrix $\mathbf{M}$ with indices in J.
$\mathbf{M}_{I,J}$	The submatrix of a matrix $\mathbf{M}$ with row indices in I and column indices in J.
$\mathbf{M}^\dagger$	The Moore-Penrose inverse of a matrix $\mathbf{M}$.
$\mathrm{diag}(\mathbf{M})$	The vector consisting of the diagonal entries of a matrix $\mathbf{M}$.
$\mathrm{trace}(\mathbf{M})$	The trace of a square matrix $\mathbf{M}$, the sum of its diagonal entries.
$\det(\mathbf{M})$	The determinant of a square matrix $\mathbf{M}$.
$\mathrm{span}\{\mathbf{M}\}$	The span of the columns of a matrix $\mathbf{M}$.
$\nabla F(x)$	The gradient of a function F at x (by default, with respect to the Euclidean inner product).
$\mathbb{E}(X)$	The expectation of a random variable X.

List of abbreviations

BI	Best improvement
CSP	Column-sampling problem
DPP	Determinantal point process
FW	Frank-Wolfe
GD	Gradient descent
HS	Hilbert-Schmidt
i.i.d.	Independent and identically distributed
ONB	Orthonormal basis
PSD	Positive semidefinite
QP	Quadratic program
RKHS	Reproducing kernel Hilbert space
SGD	Stochastic gradient descent
SPSD	Symmetric positive semidefinite
SVD	Singular value decomposition
WO	Weight optimisation

Chapter 1

Introduction

In this chapter, we provide an overview of the motivation behind the thesis, describe its main contributions, and outline the organisation of the manuscript.

1.1 Motivation

Positive semidefinite (PSD) matrices are ubiquitous in mathematics and its applications. For instance, in probability and statistics, covariance matrices are PSD matrices; such matrices also play a central role in kernel methods, an important class of techniques in machine learning and approximation theory. PSD matrices also characterise quadratic forms, and are therefore of importance in geometry and optimisation. More generally, PSD matrices correspond to discrete instances of PSD operators, and the diagonalisation of such matrices is at the core of numerous spectral approximation techniques.

From a numerical standpoint, the worst-case time complexity of diagonalising a PSD matrix of order N is cubic in N, making this operation intractable for large N (not to mention the issues related to the storage of large matrices). This has motivated the development of numerical approaches for low-rank approximation based on the notion of *column sampling*; the induced approximations are then referred to as *Nyström approximations*[1] (see for instance Williams and Seeger (2000)). The idea of approximating a PSD matrix from a sample of its columns naturally raises questions related to the characterisation of samples leading to

[1] In the literature on PSD matrix approximation, Nyström approximation refers to the low-rank approximation of PSD matrices through column sampling; although related, this terminology should not be confused with the quadrature method for the approximation of integral equations.

accurate Nyström approximations. This problem is referred to as the *column-sampling problem* (CSP). The CSP is inherently difficult; it is indeed combinatoric in nature, and in practical applications, the assessment of the quality of a given column sample is numerically challenging. As an alternative, a wide variety of heuristic-based approaches for column sampling have been developed (see Chapter 2 for an overview).

The focus of this thesis is the development and analysis of sampling strategies for the CSP which leverage connections between PSD matrices, reproducing kernel Hilbert spaces (RKHSs) and approximation in Hilbert-Schmidt (HS) spaces.

1.2 Main contributions and organisation of the manuscript

In Chapter 2, we provide an overview of the key concepts related to the approximation of general matrices via column sampling and to the assessment of the accuracy of such approximations. We then place a special emphasis on the Nyström approximation of PSD matrices, and describe some popular approaches to the CSP.

In Chapter 3, we give a detailed description of the *energy setting*, which consists of representing HS integral operators acting on an RKHS as potentials in the associated squared-kernel RKHS (see Gauthier (2024)). By interpreting PSD matrices and their approximations as HS operators, we can then relate the characterisation of low-rank approximations of PSD matrices to the approximation of specific discrete potentials.

Chapters 4 and 5 consist of the main original contributions of the manuscript. In Chapter 4, we describe a class of sequential sampling strategies for the CSP which leverage the properties of an energy-based differentiable pseudoconvex relaxation of the problem, where column samples are characterised through the non-zero entries of selection vectors; such selection vectors can be regarded as discrete measures, and together with the considered PSD matrix, define integral operators acting on the RKHS defined by the matrix (see Section 3.1.3). Following Gauthier and Suykens (2018) and Gauthier (2024), the norm of the corresponding 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 differentiable

error map on the selection-vector space (the map is in addition pseudoconvex on a specific convex cone of interest). The proposed sampling strategies relate to kernel-herding-type strategies (see e.g. Chen et al. (2010) and Bach et al. (2012)), and are based on gradient-based minimisation procedures with sparse initialisations and sparse descent directions; sparsity of the samples is then enforced via early stopping of the optimisation. Stochastic variants are also discussed, which aim at improving the computational efficiency of the approaches.

Chapter 5 focuses on the specific case of kernel matrices, where, rather than being characterised by subsets of columns, Nyström approximations can more generally be characterised by sets of landmark points. In this framework, and using a variant of the previously introduced rescaling-invariant error map, we describe a class of particle-flow-based techniques for the local optimisation of landmark points. We prove the convergence of such algorithms in the deterministic setting, and discuss their stochastic approximation.

Algorithmic implementations of the methods described in Chapters 4 and 5 are available at `https://github.com/matthutchings/energy-sampling` in the form of illustrative code examples. These examples include Python functions for the described algorithms, so that the reader may experiment with the methods using alternative data and initialisations.

Chapter 2

Low-rank approximation of positive semidefinite matrices

In this chapter, we provide an overview of the key concepts related to the approximation of general matrices via column sampling and to the assessment of the accuracy of such approximations. We then place a special emphasis on the Nyström approximation of positive semidefinite (PSD) matrices, and describe some popular approaches to the column-sampling problem (CSP).

2.1 Low-rank approximation: general case

In this section, we present some classical results concerning the low-rank approximation of general matrices (the particular case of PSD matrices will be discussed in Section 2.2).

2.1.1 Approximation accuracy

Although infinitely-many different norms could be considered, in the literature on low-rank matrix approximation, the three most commonly encountered norms used to assess the approximation accuracy are the trace, Frobenius and spectral norms; they are defined below. These norms all relate to the singular values of the considered matrices.

Definition 2.1. (Unitary matrix). With $\mathbf{I}$ denoting the $N \times N$ identity matrix, a square complex matrix $\mathbf{U} \in \mathbb{C}^{N \times N}$ is called *unitary* if $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{I}$, that is, if its inverse exists and is equal to its conjugate transpose.

Definition 2.2. (Singular value decomposition). Let $\mathbf{A} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$ be a complex matrix. A *singular value decomposition* (SVD) of $\mathbf{A}$ is a factorisation of the form

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*,$$

where $\mathbf{U} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{r}}$ and $\mathbf{V} \in \mathbb{C}^{N_\mathrm{c} \times N_\mathrm{c}}$ are unitary matrices, and $\mathbf{\Sigma} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$ is a matrix whose diagonal entries are non-negative real numbers (referred to as the *singular values* of $\mathbf{A}$) and whose off-diagonal entries are all zero. We refer to the columns of $\mathbf{U}$ and $\mathbf{V}$ as *left-* and *right-singular vectors* of $\mathbf{A}$, respectively.

Remark 2.1. The singular values of a matrix $\mathbf{A} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$ are unique up to reordering; see e.g. Trefethen and Bau (2022), Thm. 4.1 for a proof. However, the SVD of $\mathbf{A}$ is not unique, since the singular values of $\mathbf{A}$ and their corresponding left- and right-singular vectors may always be reordered, and the result is also an SVD of $\mathbf{A}$. Additionally, if some of the singular values of $\mathbf{A}$ are repeated, then another SVD of $\mathbf{A}$ can be obtained by permuting singular vectors corresponding to the same singular value. An SVD in which the singular values appear in decreasing order along the diagonal of $\mathbf{\Sigma}$ is sometimes referred to as an *ordered SVD*. ◁

We now define the trace, Frobenius and spectral norms of a matrix.

Definition 2.3. Let $\mathbf{A} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$ be a complex matrix with (i, j) entry denoted by $A_{i,j}$. Let $\sigma_1 \geqslant \ldots \geqslant \sigma_R$ denote the non-zero singular values of $\mathbf{A}$, repeated with multiplicity.

(i) The *trace norm* of $\mathbf{A}$ is given by

$$\|\mathbf{A}\|_\mathrm{tr} = \sum_{r=1}^{R} \sigma_r.$$

(ii) The *Frobenius norm* of $\mathbf{A}$ is given by

$$\|\mathbf{A}\|_\mathrm{F} = \sqrt{\sum_{i=1}^{N_\mathrm{r}} \sum_{j=1}^{N_\mathrm{c}} A_{i,j}^2} = \sqrt{\mathrm{trace}(\mathbf{A}^*\mathbf{A})} = \sqrt{\sum_{i=1}^{R} \sigma_i^2}.$$

(iii) The *spectral norm* of $\mathbf{A}$ is given by

$$\|\mathbf{A}\|_\mathrm{sp} = \sigma_1,$$

the largest singular value of $\mathbf{A}$.

For a matrix $\mathbf{A} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$, let $\hat{\mathbf{A}} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$ be an approximation of $\mathbf{A}$. The accuracy of $\hat{\mathbf{A}}$ relative to $\mathbf{A}$ is classically assessed via the trace, Frobenius or spectral norm of the approximation error, that is

$$\|\mathbf{A} - \hat{\mathbf{A}}\|_{\mathrm{tr}}, \quad \|\mathbf{A} - \hat{\mathbf{A}}\|_{\mathrm{F}}, \quad \text{or} \quad \|\mathbf{A} - \hat{\mathbf{A}}\|_{\mathrm{sp}}. \tag{2.1}$$

The closer a norm error is to zero, the more accurate the approximation $\hat{\mathbf{A}}$ is deemed to be with respect to the associated norm.

2.1.2 Rank-optimal approximations

Given a matrix norm $\|\cdot\|$ and a target matrix $\mathbf{A} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$, a *rank-optimal approximation* of $\mathbf{A}$ is a matrix $\hat{\mathbf{A}} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$ that achieves the minimum approximation error with respect to $\|\cdot\|$ among all $N_\mathrm{r} \times N_\mathrm{c}$ complex matrices of the same rank; that is, if $\mathrm{rank}(\hat{\mathbf{A}}) = m$, we have

$$\|\mathbf{A} - \hat{\mathbf{A}}\| = \min_{\mathrm{rank}(\mathbf{M})=m} \|\mathbf{A} - \mathbf{M}\|.$$

In this section, we will show that the optimal rank-m approximation of a $\mathbf{A}$ with respect to the trace, Frobenius and spectral norm errors can be obtained by truncating an SVD of $\mathbf{A}$.

Unitarily invariant norms are a class of matrix norms which are invariant with respect to left- and right-multiplication by unitary matrices.

Definition 2.4. (Unitarily invariant norm). Let $\|\cdot\|$ denote a matrix norm on $\mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$; we say that $\|\cdot\|$ is *unitarily invariant* if for all $\mathbf{A} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$ and for all unitary matrices $\mathbf{X} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{r}}$ and $\mathbf{Y} \in \mathbb{C}^{N_\mathrm{c} \times N_\mathrm{c}}$, we have

$$\|\mathbf{XA}\| = \|\mathbf{AY}\| = \|\mathbf{A}\|.$$

Proposition 2.1. The trace, Frobenius and spectral norms on $\mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$ are unitarily invariant.

Proof. Let $\mathbf{A} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{c}}$, and let $\mathbf{X} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{r}}$ and $\mathbf{Y} \in \mathbb{C}^{N_\mathrm{c} \times N_\mathrm{c}}$ be unitary. Let $\mathbf{U\Sigma V}^*$ be an SVD of $\mathbf{A}$, with $\mathbf{U} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{r}}$ and $\mathbf{V} \in \mathbb{C}^{N_\mathrm{c} \times N_\mathrm{c}}$ unitary. We have that $\mathbf{XA} = (\mathbf{XU})\mathbf{\Sigma V}^*$, and $\mathbf{AY} = \mathbf{U\Sigma V}^*\mathbf{Y} = \mathbf{U\Sigma}(\mathbf{Y}^*\mathbf{V})^*$. As products of unitary matrices, the matrices $\mathbf{XU}$ and $\mathbf{Y}^*\mathbf{V}$ are also unitary. Thus the decompositions above are SVDs of $\mathbf{XA}$ and $\mathbf{AY}$, respectively, and so the singular values of $\mathbf{XA}$ and $\mathbf{AY}$ are the same as those of $\mathbf{A}$. The result follows from Definition 2.3. □

The following theorem is a classical result which characterises the optimal rank-m approximation of a complex matrix with respect to a unitarily invariant norm.

Theorem 2.1. (Eckhart–Young–Mirsky theorem; Mirsky (1960), Thm. 3). Let $\|\cdot\|$ denote any unitarily invariant norm on $\mathbb{C}^{N_\mathrm{r}\times N_\mathrm{c}}$. Let $\mathbf{A}\in\mathbb{C}^{N_\mathrm{r}\times N_\mathrm{c}}$, and let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be a singular value decomposition of $\mathbf{A}$; write $\mathbf{U}=[\mathbf{u}_1|\cdots|\mathbf{u}_{N_\mathrm{r}}]$ and $\mathbf{V}=[\mathbf{v}_1|\cdots|\mathbf{v}_{N_\mathrm{c}}]$, where $\mathbf{u}_j,\mathbf{v}_j\in\mathbb{C}^N$ denote the j-th columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. Let $m\leqslant\min\{N_\mathrm{r},N_\mathrm{c}\}$, and consider the matrix

$$\mathbf{A}_m^\star=\sum_{k=1}^{m}\sigma_k\mathbf{u}_k\mathbf{v}_k^*, \tag{2.2}$$

obtained by truncating the SVD of $\mathbf{A}$ to its first m terms. If $m\leqslant\mathrm{rank}(\mathbf{A})$, we have

$$\inf_{\mathrm{rank}(\mathbf{M})=m}\|\mathbf{A}-\mathbf{M}\|=\|\mathbf{A}-\mathbf{A}_m^\star\|;$$

otherwise, we have $\mathbf{A}_m^\star=\mathbf{A}$.

The Eckhart–Young–Mirsky theorem combined with Proposition 2.1 tells us that an approximation $\mathbf{A}_m^\star$ of the form (2.2) is rank-optimal with respect to the trace, Frobenius and spectral norms on $\mathbb{C}^{N_\mathrm{r}\times N_\mathrm{c}}$. Due to their reliance on an SVD of the target matrix, rank-optimal approximations are often too costly to compute in applications involving large matrices, hence motivating the development of alternative low-rank approximation techniques based on column sampling.

2.1.3 Low-rank approximation via column sampling

Let $\mathbf{A}\in\mathbb{C}^{N_\mathrm{r}\times N_\mathrm{c}}$ be a complex rectangular matrix, and let the columns of $\mathbf{A}$ be indexed by $[N_\mathrm{c}]=\{1,\ldots,N_\mathrm{c}\}$. For a subset $I\subseteq[N_\mathrm{c}]$ with $|I|=m$, a low-rank approximation of $\mathbf{A}$ based on the columns of $\mathbf{A}$ with indices in I is of the form $\mathbf{A}_{\bullet,I}\mathbf{M}$ for some $\mathbf{M}\in\mathbb{C}^{m\times N_\mathrm{c}}$, where $\mathbf{A}_{\bullet,I}$ is the submatrix of $\mathbf{A}$ consisting of the columns indexed by I. A popular choice for $\mathbf{M}$ is the matrix $(\mathbf{A}_{\bullet,I})^\dagger\mathbf{A}$, where $(\mathbf{A}_{\bullet,I})^\dagger$ denotes the Moore–Penrose pseudoinverse of $\mathbf{A}_{\bullet,I}$ (see Penrose (1955), and also Barata and Hussein (2012) for a review). This is due to the fact that taking $\mathbf{M}=(\mathbf{A}_{\bullet,I})^\dagger\mathbf{A}$ gives the best approximation of $\mathbf{A}$ based on the columns indexed

by I in terms of the Frobenius norm error, that is,

$$\|\mathbf{A} - \mathbf{A}_{\bullet,I}(\mathbf{A}_{\bullet,I})^{\dagger}\mathbf{A}\|_{\mathrm{F}} = \min_{\mathbf{X}\in\mathbb{C}^{m\times N_{\mathrm{c}}}} \|\mathbf{A} - \mathbf{A}_{\bullet,I}\mathbf{X}\|_{\mathrm{F}}; \tag{2.3}$$

see the work of Drineas et al. (2008) for more details.

The search for accurate low-rank approximations of rectangular matrices through subsets of their columns is often referred to as the *column subset selection problem* (CSSP); see e.g. Boutsidis et al. (2007), Farahat et al. (2015) and Derezinski et al. (2020) for examples of contemporary approaches to the CSSP.

2.2 Positive-semidefinite matrices

We now discuss the particular case of PSD matrices, the primary focus of this thesis. The following Remark 2.2 gives expressions for the norm errors of rank-optimal approximations of a PSD matrix (with respect to the trace, Frobenius and spectral norms) in terms of its eigenvalues.

Remark 2.2. For a complex PSD matrix $\mathbf{K} \in \mathbb{C}^{N\times N}$, let $\mathbf{K}^{\star}_{m}$ be the optimal rank-m approximation of $\mathbf{K}$ of the form (2.2) for some $m \leqslant N$. The trace, Frobenius and spectral norms of the approximation error $\mathbf{K} - \mathbf{K}^{\star}_{m}$ are given by:

$$\|\mathbf{K} - \mathbf{K}^{\star}_{m}\|_{\mathrm{tr}} = \sum_{k=m+1}^{N} \lambda_k; \;\; \|\mathbf{K} - \mathbf{K}^{\star}_{m}\|_{\mathrm{F}} = \sqrt{\sum_{k=m+1}^{N} \lambda_k^2}; \;\; \|\mathbf{K} - \mathbf{K}^{\star}_{m}\|_{\mathrm{sp}} = \lambda_{m+1}, \tag{2.4}$$

where $\lambda_1 \geqslant \ldots \geqslant \lambda_N$ denote the eigenvalues of $\mathbf{K}$ (since $\mathbf{K}$ is PSD, its eigenvalues and its singular values coincide). For $m \geqslant \mathrm{rank}(\mathbf{K})$, we note that $\mathbf{K}^{\star}_{m} = \mathbf{K}$, in which case the above error norms vanish. ◁

2.2.1 Nyström approximation

As discussed in Section 2.1.3, low-rank approximations of matrices can be defined through subsets of their columns. The Nyström method is a specific instance of this type of approach for PSD matrices.

Let $\mathbf{K} \in \mathbb{C}^{N\times N}$ be a PSD matrix, and let the columns of $\mathbf{K}$ be indexed by $[N] = \{1, \ldots, N\}$; the *Nyström approximation of* $\mathbf{K}$ *induced by* $I \subseteq [N]$ is the matrix $\hat{\mathbf{K}}(I)$ defined as

$$\hat{\mathbf{K}}(I) = \mathbf{K}_{\bullet,I}(\mathbf{K}_{I,I})^{\dagger}\mathbf{K}_{I,\bullet}, \tag{2.5}$$

where $\mathbf{K}_{\bullet,I}$ denotes the submatrix of $\mathbf{K}$ consisting of the columns indexed by I, $\mathbf{K}_{I,\bullet} = (\mathbf{K}_{\bullet,I})^*$, and $\mathbf{K}_{I,I}$ denotes the principal submatrix of $\mathbf{K}$ consisting of the rows and columns indexed by I.

We refer to an index set I with $|I| = m$ as a *Nyström sample* of $\mathbf{K}$ of size m. The Nyström approximation matrix $\hat{\mathbf{K}}(I)$ is of rank at most m, and is guaranteed to be of rank m when $\mathbf{K}$ has full rank. Thus, choosing a Nyström sample of a small enough size ensures that the induced Nyström approximation $\hat{\mathbf{K}}(I)$ is low-rank.

Remark 2.3. The Nyström method is just one approach to the low-rank approximation of PSD matrices; other approaches exist in the literature, such as the work by on the sparse approximation of the Cholesky inverse via Kullback-Leibler divergence. For kernel matrices specifically, examples include approximation via random projections of the input data onto low-dimensional subspaces (see e.g. the work by Blum (2005)), and the random Fourier features approach of Rahimi and Recht (2007). An advantage of the Nyström method is that it only requires a sample of columns from $\mathbf{K}$ to build the approximation; however, this can be a limitation when compared to other contempory methods that allow for more freedom in the sampling process.

The motivations behind the Nyström approximation $\hat{\mathbf{K}}(I)$ and its properties are discussed further in Chapter 3; see also the following Remark 2.4.

Remark 2.4. A similar result to (2.3) holds for Nyström approximations of PSD matrices. For a given subset of columns I of $\mathbf{K}$, with $|I| = m \leqslant N$, the $m \times N$ matrix $\mathbf{X}_{\text{opt}} = (\mathbf{K}_{I,I})^{\dagger}\mathbf{K}_{I,\bullet}$ satisfies

$$\mathbf{X}_{\text{opt}} \in \arg\min_{\mathbf{X}\in\mathbb{C}^{m\times N}} \|\mathbf{K} - \mathbf{K}_{\bullet,I}\mathbf{X}\|_{\text{tr}}$$

see Rasmussen and Williams (2006), Chap. 8. The matrix $\mathbf{K}_{\bullet,I}\mathbf{X}_{\text{opt}}$ is precisely the Nyström approximation $\hat{\mathbf{K}}(I)$ defined in (2.5). ◁

2.2.2 The column-sampling problem

The definition of Nyström approximations naturally raises questions related to the characterisation of subsets of columns leading to accurate low-rank approximations

of a PSD matrix; we refer to the search for such subsets as the *column-sampling problem* (CSP).

The CSP is inherently difficult for two main reasons. Firstly, enumerating over all possible column samples of size m is a combinatorial problem, with $\binom{N}{m}$ different samples to consider, and secondly, the computation of the approximation norm errors becomes expensive when N is large, especially for the Frobenius and spectral norms. Indeed, the evaluation of each of the norm errors first requires the pseudoinversion of the submatrix $\mathbf{K}_{I,I}$ in (2.5), requiring $\mathcal{O}(m^3)$ operations. After this, the trace norm error is the cheapest to obtain of the three, as it only requires the computation of the N diagonal entries of the matrices $\mathbf{K}$ and $\hat{\mathbf{K}}(I)$. For the Frobenius norm error, one must compute all N^2 entries of the two matrices, and the spectral norm error additionally requires the diagonalisation of the error matrix $\mathbf{K} - \hat{\mathbf{K}}(I)$. Table 2.1 summarises the computational costs of evaluating each of these quantities.

Table 2.1: Computational complexities of evaluating the trace, Frobenius and spectral norms of the approximation error for a Nyström approximation $\hat{\mathbf{K}}(I)$ with $|I| = m \leqslant N$.

Approx. norm error	Complexity
$\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{tr}}$	$\mathcal{O}(m^3 + m^2 N)$
$\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{F}}$	$\mathcal{O}(m^3 + mN^2)$
$\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{sp}}$	$\mathcal{O}(m^3 + mN^2 + N^3)$

As a result, the direct minimisation of the trace, Frobenius or spectral norm errors over all possible Nyström samples of a given size is generally impractical. This has motivated the development and study of a wide range of heuristic-based sampling strategies for the CSP.

2.2.3 Approaches to the column-sampling problem

In this section, we discuss some popular types of sampling strategy for Nyström approximations; many other approaches exist in the literature (see e.g. Sun et al. (2015) for a review).

2.2.3.1 Naive sampling strategies

We present two examples of "naive" column-sampling strategies, that is, strategies which require a minimal amount of computation. Although enjoying interesting

properties, they typically require relatively large column samples to achieve given target accuracies when compared to more sophisticated approaches see e.g. the experiments of Gittens and Mahoney (2016).

Uniform-random sampling. Perhaps the simplest and most computationally efficient sampling strategy for the CSP involves sampling columns of the PSD matrix $\mathbf{K}$ uniformly at random without replacement. Indeed, this was the sampling strategy used in the work of Williams and Seeger (2000), the original paper on the Nyström method. Probabilistic bounds on the accuracy of uniform-sampling-induced Nyström approximations with respect to the Frobenius and spectral norm errors were obtained in Kumar et al. (2012).

Squared-diagonal-random sampling. In the work of Drineas and Mahoney (2005), the authors propose sampling columns with probabilities proportional to the squared diagonal entries of $\mathbf{K}$, that is, with probabilities

$$p_i = \frac{\mathbf{K}_{i,i}^2}{\sum_{j=1}^N \mathbf{K}_{j,j}^2}, \quad i \in [N];$$

the main result of the paper is a set of statistical bounds on the induced Frobenius and spectral Nyström approximation errors for this sampling strategy.

2.2.3.2 Leverage scores

Leverage-score-based column sampling typically involves sampling columns of the target PSD matrix $\mathbf{K}$ with probabilities proportional to their statistical leverage scores. These scores are interpreted as measures of the relative "importance" of the columns of $\mathbf{K}$.

Definition 2.5. (Leverage scores; Gittens and Mahoney (2016)). Let $N \in \mathbb{N}$, and let $\mathbf{K} \in \mathbb{C}^{N \times N}$ be a complex PSD matrix with eigenvalue decomposition $\mathbf{K} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^*$, where $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$ is the diagonal matrix formed from the eigenvalues of $\mathbf{K}$ in descending order, repeated with multiplicity, and the columns of $\mathbf{P} \in \mathbb{C}^{N \times N}$ consist of the corresponding eigenvectors. Let $k \leqslant N$, and let $\mathbf{P}_k \in \mathbb{C}^{N \times k}$ be the matrix formed from the first k columns of $\mathbf{P}$, that is, the top k eigenvectors of $\mathbf{K}$.

For $j \in [N]$, the j-th *leverage score* of $\mathbf{K}$ relative to its optimal rank-k approximation is given by

$$l_j^k = \|(\mathbf{P}_k)_{j,\bullet}\|^2,$$

where $(\mathbf{P}_k)_{j,\bullet}$ denotes the j-th row of $\mathbf{P}_k$.

A number of leverage-score-based approaches to the CSP can be found in the work of Gittens and Mahoney (2016), where the authors describe column-sampling techniques based on both exact and approximate leverage scores. For the approximate leverage score case, the authors use an algorithm from the paper (Drineas et al., 2012) which computes approximations of the leverage scores using a modified Johnson–Lindenstrauss Transform first proposed in the work (Ailon and Chazelle, 2006). Leverage scores may also be approximated more naively through uniform sampling; see e.g. Cohen et al. (2015).

More recently, there has been interest in the use of so-called ridge leverage scores and their approximations for solving the CSP. Ridge leverage scores add an additional regularisation parameter $\lambda > 0$ which is often problem-dependent; techniques involving ridge leverage scores have been developed to approximate kernel ridge regression problems, see, for example, the works of Alaoui and Mahoney (2015) and Chen and Yang (2021).

Definition 2.6. (Ridge leverage scores; Alaoui and Mahoney (2015)). Let $\mathbf{K} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^* \in \mathbb{C}^{N\times N}$ be a complex PSD matrix, and let $\lambda > 0$. Let the singular values of $\mathbf{K}$, repeated with multiplicity, be denoted by $\sigma_1 \geqslant \ldots \geqslant \sigma_N \geqslant 0$. For $j \in [N]$, the j-th λ-*ridge leverage score* of $\mathbf{K}$ is given by

$$l_j(\lambda) = \sum_{i=1}^{N} \frac{\sigma_i}{\sigma_i + N\lambda} P_{i,j}^2.$$

As with the standard leverage scores, ridge leverage scores are costly to obtain due to the need to diagonalise $\mathbf{K}$. Techniques involving approximate ridge leverage scores have been developed to address this issue, two of which being the Recursive-RLS algorithm of Musco and Musco (2017) and the Divide and Conquer algorithm proposed in Cherfaoui et al. (2022).

2.2.3.3 Determinantal point processes

With origins in the field of random matrix theory (see e.g. Mehta and Gaudin (1960) and Ginibre (1965)), determinantal point processes (DPPs) were formalised in the work of Macchi (1975), albeit under the name of "fermion processes"; they describe a class of point processes which exhibit repulsion between points. DPPs also appear naturally in the field of randomised numerical linear algebra (see e.g. Derezinski and Mahoney (2021)), and have been utilised in a variety of machine learning applications due to their ability to sample data in a diverse way (see Kulesza et al. (2012) for a review).

Definition 2.7. (DPP; Kulesza et al. (2012)). Let $N \in \mathbb{N}$, and let $\mathbf{K} \in \mathbb{R}^{N \times N}$ be a real symmetric positive semidefinite (SPSD) matrix which satisfies $0 \preceq \mathbf{K} \preceq \mathbf{I}$, with $\mathbf{I}$ the $N \times N$ identity matrix, and where $\preceq$ denotes the Löwner partial ordering (see e.g. Zhan (2004), Chap. 1).

A *determinantal point process* (DPP) on $[N]$ with marginal kernel $\mathbf{K}$ is a probability measure $\mathbb{P}$ on the power set $2^{[N]}$ which satisfies the following: if $S \subseteq [N]$ is a random sample from $\mathbb{P}$, then for every $T \subseteq [N]$, we have

$$\mathbb{P}(\{T \subseteq S\}) = \det(\mathbf{K}_{T,T});$$

we say that $S \sim \mathrm{DPP}(\mathbf{K})$.

A particular type of DPP is the $\mathbf{L}$-ensemble, which removes the restriction $0 \preceq \mathbf{K} \preceq \mathbf{I}$ in Definition 2.7.

Definition 2.8. ($\mathbf{L}$-ensemble; Kulesza et al. (2012)). Let $\mathbf{L} \in \mathbb{R}^{N \times N}$ be a real SPSD matrix. An $\mathbf{L}$-ensemble is a probability measure $\mathbb{P}_{\mathbf{L}}$ on $2^{[N]}$ which satisfies the following: if $S \subseteq [N]$ is a random sample from $\mathbb{P}_{\mathbf{L}}$, then for every $T \subseteq [N]$, we have

$$\mathbb{P}_{\mathbf{L}}(\{S = T\}) \propto \det(\mathbf{L}_{T,T}).$$

We note that every $\mathbf{L}$-ensemble is a DPP with marginal kernel $\mathbf{K} = \mathbf{L}(\mathbf{L} + \mathbf{I})^{-1}$.

For the CSP, it is often desirable to sample subsets of a fixed size k. The result is a k-DPP, defined below.

Definition 2.9. (k-DPP; Kulesza et al. (2012)). A *k-DPP* is an **L**-ensemble conditioned on subsets of size k. If $\mathbb{P}^k_{\mathbf{L}}$ is a k-DPP, and S is a random sample from $\mathbb{P}^k_{\mathbf{L}}$, then for every $T \subseteq [N]$ with $|T| = k$, we have

$$\mathbb{P}^k_{\mathbf{L}}(S = T) = \frac{\det(\mathbf{L}_{T,T})}{\sum_{|T'|=k} \det(\mathbf{L}_{T',T'})}.$$

For rectangular matrices and the CSSP (see Section 2.1.3), the authors of Derezinski et al. (2020) use k-DPP sampling to build Nyström approximations of a PSD matrix, giving theoretical guarantees on the resulting trace approximation errors. There have also been efforts to improve the computational efficiency of DPP sampling, including, but not limited to, the Gibbs sampler described in the work of Li et al. (2016) and the DPP-VFX sampler of Derezinski et al. (2019).

Chapter 3

Reproducing kernel Hilbert spaces and kernel energies

In this chapter, we describe the link between the low-rank approximation of positive semidefinite (PSD) matrices, the approximation of Hilbert–Schmidt (HS) integral operators with PSD kernels, and the approximation of potentials in reproducing kernel Hilbert spaces (RKHSs) with squared kernels. Section 3.1 is devoted to general results about RKHSs, HS operators and the notions of energies and potentials in RKHSs; the connection between this setting and the low-rank approximation of PSD matrices is then investigated in Section 3.2.

3.1 Hilbert–Schmidt operators and reproducing kernel Hilbert spaces

In this section, we give a brief overview of some key concepts in Hilbertian analysis. Notably, we describe the connections between the approximation of integral operators on RKHSs and the approximation of potentials, which is central to our study.

3.1.1 Reproducing kernel Hilbert spaces

In this section and throughout this chapter, we consider reproducing kernel Hilbert spaces of $\mathbb{C}$-valued functions for generality, however the results also hold for RKHSs of real-valued functions. We assume that all inner products are linear in the second argument and conjugate-linear in the first.

Definition 3.1. (RKHS). Let $\mathscr{X}$ be a general set. A *reproducing kernel Hilbert space* of $\mathbb{C}$-valued functions on $\mathscr{X}$ is a Hilbert space $\mathcal{H}$ of functions from $\mathscr{X}$ to $\mathbb{C}$ such that for every $x \in \mathscr{X}$, the evaluation functional E_x, defined by $E_x[f] = f(x)$, $f \in \mathcal{H}$, is bounded.

Let $\mathcal{H}$ be an RKHS of $\mathbb{C}$-valued functions on $\mathscr{X}$; by the Riesz representation theorem (see e.g. Roman et al. (2005), Thm. 13.32), for every $x \in \mathscr{X}$, there exists a unique $k_x \in \mathcal{H}$ such that

$$f(x) = E_x[f] = \langle k_x \,|\, f \rangle_{\mathcal{H}}, \quad f \in \mathcal{H}, \tag{3.1}$$

where $\langle \cdot \,|\, \cdot \rangle_{\mathcal{H}}$ denotes the inner product on $\mathcal{H}$. The function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$ defined by

$$K(x, y) = k_y(x) = \langle k_x \,|\, k_y \rangle_{\mathcal{H}}, \quad x, y \in \mathscr{X}, \tag{3.2}$$

is called the *reproducing kernel for* $\mathcal{H}$.

Remark 3.1. For any RKHS $\mathcal{H}$, its associated reproducing kernel K is PSD; indeed, for $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathscr{X}$ and $c_1, \ldots, c_n \in \mathbb{C}$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \overline{c_i} c_j K(x_i, x_j) = \left\langle \sum_{i=1}^{n} c_i k_{x_i} \,\middle|\, \sum_{j=1}^{n} c_j k_{x_j} \right\rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} c_i k_{x_i} \right\|_{\mathcal{H}}^2 \geq 0.$$

Conversely, the Moore–Aronszajn theorem (see Aronszajn (1950), pg. 344) states that any PSD kernel function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$ defines an RKHS on $\mathscr{X}$ for which it is the reproducing kernel. ◁

3.1.2 Hilbert–Schmidt operators

This brief section introduces the notion of Hilbert spaces of Hilbert–Schmidt operators between two Hilbert spaces (see e.g. Dunford and Schwartz (1975) for more details).

Definition 3.2. (HS operator). Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces over $\mathbb{C}$, equipped with the norms $\| \cdot \|_{\mathcal{H}_1}$ and $\| \cdot \|_{\mathcal{H}_2}$, respectively. Let $A : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator. A is called *Hilbert–Schmidt* (HS) if there exists an orthonormal basis (ONB) $\{h_j\}_{j \in J}$ of $\mathcal{H}_1$ such that

$$\sum_{j \in J} \|A h_j\|_{\mathcal{H}_2}^2 < +\infty.$$

Let A and B be two HS operators from $\mathcal{H}_1$ to $\mathcal{H}_2$, and let $\{h_j\}_{j\in J}$ be an ONB of $\mathcal{H}_1$. We define

$$\langle A \,|\, B\rangle_{\mathrm{HS}(\mathcal{H}_1,\mathcal{H}_2)} = \sum_{j\in J} \langle Ah_j \,|\, Bh_j\rangle_{\mathcal{H}_2};$$

notably, $\langle A \,|\, B\rangle_{\mathrm{HS}(\mathcal{H}_1,\mathcal{H}_2)}$ does not depend on the choice of the considered ONB (see Dunford and Schwartz (1975), Chap. XI Sec. 6).

Endowed with the Hermitian form $\langle \cdot \,|\, \cdot \rangle_{\mathrm{HS}(\mathcal{H}_1,\mathcal{H}_2)}$, the linear space $\mathrm{HS}(\mathcal{H}_1,\mathcal{H}_2)$ of all HS operators from $\mathcal{H}_1$ to $\mathcal{H}_2$ is a Hilbert space. The corresponding HS norm is defined as

$$\|A\|_{\mathrm{HS}(\mathcal{H}_1,\mathcal{H}_2)} = \sqrt{\langle A \,|\, A\rangle_{\mathrm{HS}(\mathcal{H}_1,\mathcal{H}_2)}} = \sqrt{\sum_{j\in J} \|Ah_j\|_{\mathcal{H}_2}^2}, \quad A \in \mathrm{HS}(\mathcal{H}_1,\mathcal{H}_2).$$

Remark 3.2. In the framework of Definition 3.2, in the case $\mathcal{H}_2 = \mathcal{H}_1$, we refer to A as a HS operator on $\mathcal{H}_1$ and write $A \in \mathrm{HS}(\mathcal{H}_1)$. ◁

3.1.3 Integral operators and the energy setting

In this section, we show how HS integral operators involving a PSD kernel can be interpreted as potentials in the associated squared-kernel RKHS; we refer to this as the *energy setting*, and this is the framework in which the approaches to the CSP described in Chapters 4 and 5 are based.

Let $\mathcal{H}$ be a separable RKHS, i.e. an RKHS with a countable ONB, with associated kernel function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$. Let Σ be a σ-algebra on $\mathscr{X}$, so that $(\mathscr{X}, \Sigma)$ is a measurable space. We make the following assumptions on K and Σ:

- for all $t \in \mathscr{X}$, the function $k_t : x \mapsto K(x,t)$, $x \in \mathscr{X}$, is measurable;
- the function $x \mapsto K(x,x)$, $x \in \mathscr{X}$, is measurable.

This implies that every function $h \in \mathcal{H}$ is measurable (see Steinwart and Christmann (2008) for a detailed discussion on kernels and measurability).

Let $\mathcal{M}$ denote the set of all signed measures on $(\mathscr{X}, \Sigma)$, and let $\mathcal{M}_+$ denote the real convex cone of all nonnegative measures in $\mathcal{M}$. Define the set

$$\mathcal{T}(K) = \left\{ \mu \in \mathcal{M} \,\middle|\, \tau_\mu = \int_{\mathscr{X}} K(x,x) \mathrm{d}|\mu|(x) < +\infty \right\}, \tag{3.3}$$

where $|\mu|$ denotes the variation of $\mu \in \mathcal{M}$; denote by $\mathcal{T}_+(K) = \mathcal{T}(K) \cap \mathcal{M}_+$ the real convex cone of nonnegative measures in $\mathcal{T}(K)$.

3.1.3.1 Integral operators with PSD kernels

For $h, f \in \mathcal{H}$ and $\mu \in \mathcal{T}(K)$, consider the integral

$$\mathcal{I}_{h,f,\mu} = \int_{\mathscr{X}} \overline{h(t)} f(t) \mathrm{d}\mu(t);$$

from the Cauchy–Schwarz inequality in $\mathcal{H}$ and equation (3.1), we have

$$\begin{aligned} |\mathcal{I}_{h,f,\mu}| \leqslant \int_{\mathscr{X}} |\overline{h(t)}||f(t)|\mathrm{d}|\mu|(t) &= \int_{\mathscr{X}} |\langle k_t \,|\, h\rangle_{\mathcal{H}}||\langle k_t \,|\, f\rangle_{\mathcal{H}}|\mathrm{d}|\mu|(t) \\ &\leqslant \int_{\mathscr{X}} \|h\|_{\mathcal{H}}\|f\|_{\mathcal{H}}\|k_t\|_{\mathcal{H}}^2 \mathrm{d}|\mu|(t) \\ &= \|h\|_{\mathcal{H}}\|f\|_{\mathcal{H}} \int_{\mathscr{X}} K(t,t)\mathrm{d}|\mu|(t) \\ &= \tau_\mu \|h\|_{\mathcal{H}}\|f\|_{\mathcal{H}}, \end{aligned} \tag{3.4}$$

where τ_μ is defined in (3.3). By (3.4), the linear map $\Xi_{h,\mu} : f \mapsto \mathcal{I}_{h,f,\mu}$, $f \in \mathcal{H}$, is bounded, therefore, by the Riesz representation theorem, for all $h \in \mathcal{H}$ there exists $L_\mu[h] \in \mathcal{H}$ such that

$$\Xi_{h,\mu}[f] = \int_{\mathscr{X}} \overline{h(t)} f(t) \mathrm{d}\mu(t) = \langle L_\mu[h] \,|\, f\rangle_{\mathcal{H}}, \quad f \in \mathcal{H}. \tag{3.5}$$

Let $h \in \mathcal{H}$, and let $L_\mu[h]$ be as above. For all $x \in \mathscr{X}$, we have

$$\begin{aligned} L_\mu[h](x) = \langle k_x \,|\, L_\mu[h]\rangle &= \overline{\langle L_\mu[h] \,|\, k_x\rangle} \\ &= \overline{\int_{\mathscr{X}} \overline{h(t)} k_x(t) \mathrm{d}\mu(t)} \\ &= \int_{\mathscr{X}} \overline{k_x(t)} h(t) \mathrm{d}\mu(t) \\ &= \int_{\mathscr{X}} K(x,t) h(t) \mathrm{d}\mu(t). \end{aligned} \tag{3.6}$$

Lemma 3.1. For $\mu \in \mathcal{T}(K)$, the integral operator $L_\mu : \mathcal{H} \to \mathcal{H}$, defined by (3.6) for $x \in \mathscr{X}$ and $h \in \mathcal{H}$, is a Hilbert–Schmidt operator on $\mathcal{H}$.

Proof. Let $\{h_j\}_{j\in\mathbb{J}}$ be an ONB of $\mathcal{H}$. From (3.5) and (3.6), we have

$$\begin{aligned} \|L_\mu\|_{\mathrm{HS}(\mathcal{H})}^2 = \sum_{j\in\mathbb{J}} \|L_\mu[h_j]\|_{\mathcal{H}}^2 &= \sum_{j\in\mathbb{J}} \langle L_\mu[h_j] \,|\, L_\mu[h_j]\rangle_{\mathcal{H}} \\ &= \sum_{j\in\mathbb{J}} \int_{\mathscr{X}} \overline{h_j(x)} L_\mu[h_j](x) \mathrm{d}\mu(x) \\ &= \sum_{j\in\mathbb{J}} \iint_{\mathscr{X}} \overline{h_j(x)} K(x,t) h_j(t) \mathrm{d}\mu(t)\mathrm{d}\mu(x) \\ &= \iint_{\mathscr{X}} K(x,t) \sum_{j\in\mathbb{J}} \overline{h_j(x)} h_j(t) \mathrm{d}\mu(t)\mathrm{d}\mu(x) \end{aligned}$$

$$
\begin{aligned}
&= \iint_{\mathscr{X}} K(x,t)K(t,x)\mathrm{d}\mu(t)\mathrm{d}\mu(x) \\
&= \iint_{\mathscr{X}} |K(x,t)|^2\mathrm{d}\mu(t)\mathrm{d}\mu(x) \\
&\leqslant \iint_{\mathscr{X}} K(x,x)K(t,t)\mathrm{d}\mu(t)\mathrm{d}\mu(x) = \tau_\mu^2,
\end{aligned}
$$

where the inequality follows from (3.2) and Cauchy–Schwarz, and the exchange of the integrals and the sum is justified by Fubini's theorem. □

Remark 3.3. For $\mu \in \mathcal{T}_+(K)$, the integral operator $L_\mu \in \mathrm{HS}(\mathcal{H})$ can be written as $L_\mu = \iota_\mu^* \iota_\mu$, with $\iota_\mu : \mathcal{H} \to \mathcal{L}^2(\mu)$ the natural embedding of $\mathcal{H}$ into $\mathcal{L}^2(\mu)$, where $\iota_\mu[h]$ is the equivalence class of all measurable functions that are μ-almost everywhere equal to $h \in \mathcal{H}$. We may then also consider the operators

$$
\iota_\mu^* : \mathcal{L}^2(\mu) \to \mathcal{H}, \quad \iota_\mu \iota_\mu^* : \mathcal{L}^2(\mu) \to \mathcal{L}^2(\mu), \quad \text{and} \quad \iota_\mu \iota_\mu^* \iota_\mu : \mathcal{H} \to \mathcal{L}^2(\mu). \tag{3.7}
$$

For a detailed discussion and study of these types of integral operators in their most general setting, we refer the reader to the work of Gauthier (2024). ◁

3.1.3.2 Potentials in squared-kernel RKHSs

The function $|K|^2 = K \cdot \overline{K}$, given by $|K(x,t)|^2$, $x, t \in \mathscr{X}$, is a valid kernel function, since the product of two kernels is a kernel (see e.g. Paulsen and Raghupathi (2016), Thm. 5.24 and Cor. 5.27). Let $\mathcal{G}$ be the RKHS on $\mathscr{X}$ for which $|K|^2$ is reproducing. For $g \in \mathcal{G}$ and $\mu \in \mathcal{T}(K)$, we define

$$
\mathcal{I}_{g,\mu} = \int_{\mathscr{X}} g(t)\mathrm{d}\mu(t).
$$

Similarly to (3.4), we have

$$
|\mathcal{I}_{g,\mu}| \leqslant \int_{\mathscr{X}} |g(t)|\mathrm{d}|\mu|(t) \leqslant \tau_\mu \|g\|_{\mathcal{G}},
$$

and by the Riesz representation theorem, there exists $g_\mu \in \mathcal{G}$ such that

$$
\mathcal{I}_{g,\mu} = \langle g_\mu \,|\, g \rangle_{\mathcal{G}}, \quad g \in \mathcal{G};
$$

we refer to g_μ as the *potential* of μ in $\mathcal{G}$. For all $x \in \mathscr{X}$, we have

$$
g_\mu(x) = \langle |k_x|^2 \,|\, g_\mu \rangle = \overline{\int_{\mathscr{X}} |k_x|^2(t)\mathrm{d}\mu(t)} = \int_{\mathscr{X}} |K(x,t)|^2\mathrm{d}\mu(t),
$$

where the function $|k_x|^2 \in \mathcal{G}$ is given by $|k_x|^2(t) = |K|^2(x,t) = |K(x,t)|^2$, $t \in \mathscr{X}$. It follows that

$$\|g_\mu\|_{\mathcal{G}}^2 = \langle g_\mu \,|\, g_\mu \rangle_{\mathcal{G}} = \int_{\mathscr{X}} g_\mu(x) \mathrm{d}\mu(x) = \iint_{\mathscr{X}} |K(x,t)|^2 \mathrm{d}\mu(t) \mathrm{d}\mu(x) = \|L_\mu\|_{\mathrm{HS}(\mathcal{H})}^2,$$

and we refer to the quantity $\|g_\mu\|_{\mathcal{G}}^2$ as the *energy* of μ in $\mathcal{G}$. In this way, for every $\mu \in \mathcal{T}(K)$, the integral operator L_μ in $\mathrm{HS}(\mathcal{H})$ given by (3.6) is naturally associated with the potential $g_\mu \in \mathcal{G}$.

In particular, for μ and ν in $\mathcal{T}(K)$, by Cauchy–Schwarz, we have

$$\|L_\mu - L_\nu\|_{\mathrm{HS}(\mathcal{H})} = \|g_\mu - g_\nu\|_{\mathcal{G}} = \sup_{g \in \mathcal{B}_{\mathcal{G}}} \left| \int_{\mathscr{X}} g(t) \mathrm{d}\mu(t) - \int_{\mathscr{X}} g(t) \mathrm{d}\nu(t) \right|,$$

with $\mathcal{B}_{\mathcal{G}}$ the closed unit ball of $\mathcal{G}$. The map $(\mu, \nu) \mapsto \|L_\mu - L_\nu\|_{\mathrm{HS}(\mathcal{H})}$, $\mu, \nu \in \mathcal{T}(K)$, is therefore a generalised *maximum mean discrepancy* (MMD; see e.g. Muandet et al. (2017) and Tolstikhin et al. (2016)) for the squared-kernel function $|K|^2$ (by generalised, we mean that not only probability measures are considered). The representation of this map as the norm difference of potentials in $\mathcal{G}$ will form the basis for the energy-based optimisation framework considered in Chapters 4 and 5.

3.2 Reproducing kernel Hilbert spaces and PSD matrix approximation

The entries of a PSD matrix $\mathbf{K} \in \mathbb{C}^{N \times N}$ can be interpreted as the values of a kernel function $K : [N] \times [N] \to \mathbb{C}$, and as such, define an RKHS of $\mathbb{C}$-valued functions on $[N]$. This RKHS can be identified with the subspace $\mathcal{H} = \mathrm{span}\{\mathbf{K}\} \subseteq \mathbb{C}^N$ endowed with the inner product

$$\langle \boldsymbol{h} \,|\, \boldsymbol{f} \rangle_{\mathcal{H}} = \boldsymbol{h}^* \mathbf{K}^\dagger \boldsymbol{f}, \quad \boldsymbol{h}, \boldsymbol{f} \in \mathcal{H}. \tag{3.8}$$

A subset $I \subseteq [N]$ indexing a column sample of $\mathbf{K}$ then defines a closed linear subspace $\mathcal{H}_I = \mathrm{span}\{\mathbf{K}_{\bullet, I}\} \subseteq \mathcal{H}$ of dimension $|I|$. Introducing the matrix $\mathbf{P}_I = \mathbf{K}_{\bullet, I} (\mathbf{K}_{I,I})^\dagger \mathbf{I}_{I, \bullet} \in \mathbb{C}^{N \times N}$, we have

$$\hat{\mathbf{K}}(I) = \mathbf{P}_I \mathbf{K} = \mathbf{K} \mathbf{P}_I^* = \mathbf{P}_I \mathbf{K} \mathbf{P}_I^*,$$

where the matrix $\hat{\mathbf{K}}(I)$ is the Nyström approximation of $\mathbf{K}$ induced by I, given by (2.5). The matrix $\mathbf{P}_I$ corresponds to the orthogonal projection from $\mathcal{H}$ onto $\mathcal{H}_I$;

for all $\boldsymbol{h}, \boldsymbol{f} \in \mathcal{H}$, we indeed have

$$\operatorname{span}\{\mathbf{P}_I\mathbf{K}\} = \mathcal{H}_I, \;\; \mathbf{P}_I^2 = \mathbf{P}_I, \;\; \text{and} \;\; \langle \boldsymbol{h} \,|\, \mathbf{P}_I \boldsymbol{f} \rangle_{\mathcal{H}} = \langle \mathbf{P}_I \boldsymbol{h} \,|\, \boldsymbol{f} \rangle_{\mathcal{H}},$$

so that the matrix $\hat{\mathbf{K}}(I) = \mathbf{P}_I\mathbf{K}$ is the reproducing kernel for the subspace $\mathcal{H}_I$; see e.g. Paulsen and Raghupathi (2016).

3.2.1 Nyström approximations and projections

In this section, we discuss how orthogonal projections of a PSD matrix $\mathbf{K}$ may be viewed as Hilbert–Schmidt operators, and show that we are able to recover the trace, Frobenius and spectral norms of the approximation error in this framework.

Denoting by $\mathcal{E}$ the Euclidean Hilbert space $\mathbb{C}^N$ (with $\langle \mathbf{u} \,|\, \mathbf{v} \rangle_{\mathcal{E}} = \mathbf{u}^*\mathbf{v}$, $\mathbf{u}, \mathbf{v} \in \mathcal{E}$), 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

$$\langle \boldsymbol{h} \,|\, \mathbf{K}\mathbf{v} \rangle_{\mathcal{H}} = \langle \boldsymbol{h} \,|\, \mathbf{v} \rangle_{\mathcal{E}}, \quad \boldsymbol{h} \in \mathcal{H}, \mathbf{v} \in \mathcal{E}. \tag{3.9}$$

We denote by $\{\mathbf{e}_i\}_{i \in [N]}$ the canonical basis of $\mathbb{C}^N$. In light of (3.9), the matrices $\mathbf{K}$ and $\mathbf{P}_I\mathbf{K}$ can be regarded as HS operators from, and to, $\mathcal{E}$ or $\mathcal{H} = \operatorname{span}\{\mathbf{K}\}$.

Case 1: $\mathcal{E} \to \mathcal{H}$ Following Section 3.1.2, we set

$$\operatorname{HS}(\mathcal{E}, \mathcal{H}) = \{\mathbf{M} \in \mathbb{C}^{N \times N} \mid \operatorname{span}\{\mathbf{M}\} \subseteq \mathcal{H}\}. \tag{3.10}$$

For all $\mathbf{M}, \mathbf{T} \in \operatorname{HS}(\mathcal{E}, \mathcal{H})$, we have

$$\langle \mathbf{M} \,|\, \mathbf{T} \rangle_{\operatorname{HS}(\mathcal{E},\mathcal{H})} = \sum_{i=1}^{N} \langle \mathbf{M}\mathbf{e}_i \,|\, \mathbf{T}\mathbf{e}_i \rangle_{\mathcal{H}} = \operatorname{trace}(\mathbf{M}^*\mathbf{K}^\dagger\mathbf{T}).$$

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

The trace norm of the approximation error corresponds to the squared HS norm of the PSD error matrix $\mathbf{K} - \hat{\mathbf{K}}(I)$ when interpreted as an operator from $\mathcal{E}$ to $\mathcal{H}$; indeed, setting $\mathbf{P}_{0I} = \mathbf{I} - \mathbf{P}_I$ (so that $\mathbf{K} - \hat{\mathbf{K}}(I) = \mathbf{P}_{0I}\mathbf{K} = \mathbf{K}\mathbf{P}_{0I}^*$) and observing that the matrix $\mathbf{P}_{0I}$ corresponds to an orthogonal projection on $\mathcal{H}$, from (3.9) we obtain

$$\|\mathbf{P}_{0I}\mathbf{K}\|_{\operatorname{HS}(\mathcal{E},\mathcal{H})}^2 = \sum_{i=1}^{N} \|\mathbf{P}_{0I}\mathbf{K}\mathbf{e}_i\|_{\mathcal{H}}^2 = \sum_{i=1}^{N} \langle \mathbf{P}_{0I}\mathbf{K}\mathbf{e}_i \,|\, \mathbf{K}\mathbf{e}_i \rangle_{\mathcal{H}}$$

$$= \sum_{i=1}^{N} \langle \mathbf{P}_{0I}\mathbf{K}\mathbf{e}_i \,|\, \mathbf{e}_i \rangle_{\mathcal{E}} = \operatorname{trace}(\mathbf{P}_{0I}\mathbf{K}) = \|(\mathbf{K} - \hat{\mathbf{K}}(I))\|_{\mathrm{tr}}. \quad (3.11)$$

Case 2: $\mathcal{H} \to \mathcal{E}$. The Frobenius and spectral norms of the matrix $\mathbf{K} - \hat{\mathbf{K}}(I)$ correspond to the HS and spectral norms of this matrix when regarded as an operator on $\mathcal{E}$. For the Frobenius norm, we observe that

$$\|\mathbf{P}_{0I}\mathbf{K}\|^2_{\mathrm{HS}(\mathcal{E})} = \sum_{i=1}^{N} \|\mathbf{P}_{0I}\mathbf{K}\mathbf{e}_i\|^2_{\mathcal{E}} = \sum_{i=1}^{N} \mathbf{e}_i^* \mathbf{K}\mathbf{P}_{0I}^*\mathbf{P}_{0I}\mathbf{K}\mathbf{e}_i$$
$$= \operatorname{trace}((\mathbf{P}_{0I}\mathbf{K})^*(\mathbf{P}_{0I}\mathbf{K})) = \|\mathbf{K} - \hat{\mathbf{K}}(I)\|^2_{\mathrm{F}}. \quad (3.12)$$

Case 3: $\mathcal{H} \to \mathcal{H}$. In the spirit of (3.10), we have the following characterisation of the space $\mathrm{HS}(\mathcal{H})$:

$$\mathrm{HS}(\mathcal{H}) = \{\mathbf{M} \in \mathbb{C}^{N\times N} \mid \operatorname{span}\{\mathbf{M}\mathbf{K}\} \subseteq \mathcal{H}\}, \quad (3.13)$$

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 $\boldsymbol{h} \in \mathcal{H}$. Observe 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. Paulsen and Raghupathi (2016)); it follows that for $\mathbf{M}, \mathbf{T} \in \mathrm{HS}(\mathcal{H})$,

$$\langle \mathbf{M} \,|\, \mathbf{T} \rangle_{\mathrm{HS}(\mathcal{H})} = \sum_{j\in\mathbb{J}} \langle \mathbf{M}\boldsymbol{h}_j \,|\, \mathbf{T}\boldsymbol{h}_j \rangle_{\mathcal{H}} = \operatorname{trace}(\mathbf{K}\mathbf{M}^*\mathbf{K}^\dagger\mathbf{T}) \quad (3.14)$$

Endowed with $\langle \cdot \,|\, \cdot \rangle_{\mathrm{HS}(\mathcal{H})}$, the linear space $\mathrm{HS}(\mathcal{H})$ is a semi-Hilbert space, that is, a complete inner product space in which the inner product is only required to be PSD. We have that $\|\mathbf{M}\|_{\mathrm{HS}(\mathcal{H})} = 0$ if and only if $\mathbf{M}\mathbf{K} = 0$ (see Remark 3.4). If $\mathbf{K}$ is invertible, then $\mathrm{HS}(\mathcal{H})$ is a Hilbert space.

Remark 3.4. When the matrix $\mathbf{K}$ is singular, the matrices representing a given operator on $\mathcal{H}$ are nonunique. Indeed, for $\mathbf{v} \in \mathbb{C}^N$ with $\mathbf{v} \neq 0$ and $\mathbf{K}\mathbf{v} = 0$, we have $\mathbf{v}^*\boldsymbol{h} = 0$ for all $\boldsymbol{h} \in \mathcal{H}$; 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 $\mathcal{H}$. ◁

We now present a result which provides a link between $\|\cdot\|_{\mathrm{HS}(\mathcal{H})}$ and the Frobenius inner product on $\mathcal{E}$ for orthogonal projections of $\mathbf{K}$.

Lemma 3.2. Let $\mathbf{P}, \mathbf{Q} \in \mathbb{C}^{N\times N}$ be two matrices corresponding to orthogonal projections onto closed linear subspaces of $\mathcal{H}$. We have

$$\|\mathbf{P}\mathbf{K}\mathbf{Q}\|^2_{\mathrm{HS}(\mathcal{H})} = \langle \mathbf{P}\mathbf{K} \,|\, \mathbf{Q}\mathbf{K} \rangle_{\mathrm{F}}.$$

Proof. We first observe that $\mathbf{PK} = \mathbf{KP}^* = \mathbf{PKP}^*$ (the same property holds for $\mathbf{Q}$). From (3.9), we indeed have

$$\begin{aligned}\mathbf{e}_i^*\mathbf{PKe}_j &= \mathbf{e}_i^*\mathbf{KK}^\dagger\mathbf{PKe}_j = \langle \mathbf{Ke}_i \,|\, \mathbf{PKe}_j\rangle_{\mathcal{H}} \\ &= \langle \mathbf{PKe}_i \,|\, \mathbf{Ke}_j\rangle_{\mathcal{H}} = \langle \mathbf{PKe}_i \,|\, \mathbf{e}_j\rangle_{\mathcal{E}} = \langle \mathbf{Ke}_i \,|\, \mathbf{P}^*\mathbf{e}_j\rangle_{\mathcal{E}} \\ &= \langle \mathbf{Ke}_i \,|\, \mathbf{KP}^*\mathbf{e}_j\rangle_{\mathcal{H}} = \mathbf{e}_i^*\mathbf{KK}^\dagger\mathbf{KP}^*\mathbf{e}_j = \mathbf{e}_i^*\mathbf{KP}^*\mathbf{e}_j, \quad i,j \in [N];\end{aligned}$$

in particular, the equality $\mathbf{e}_i^*\mathbf{PKe}_j = \mathbf{e}_i^*\mathbf{KK}^\dagger\mathbf{PKe}_j$ follows by noticing that since $\mathbf{PKe}_j \in \mathcal{H}$, there exists $\boldsymbol{\alpha} \in \mathbb{C}^N$ such that $\mathbf{PKe}_j = \mathbf{K}\boldsymbol{\alpha}$. From (3.14), we then get

$$\begin{aligned}\|\mathbf{PKQ}\|_{\mathrm{HS}(\mathcal{H})}^2 &= \mathrm{trace}(\mathbf{KQ}^*\mathbf{KP}^*\mathbf{K}^\dagger\mathbf{PKQ}) = \mathrm{trace}(\mathbf{KQ}^*\mathbf{PKK}^\dagger\mathbf{KP}^*\mathbf{Q}) \\ &= \mathrm{trace}(\mathbf{PKP}^*\mathbf{QKQ}^*) = \mathrm{trace}(\mathbf{KP}^*\mathbf{QK}) = \langle \mathbf{PK} \,|\, \mathbf{QK}\rangle_{\mathrm{F}},\end{aligned}$$

where we have used the fact that $\mathrm{trace}(\mathbf{AB}) = \mathrm{trace}(\mathbf{BA})$ for all $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{N\times N}$. □

Applying Lemma 3.2 with $\mathbf{P} = \mathbf{P}_{0I}$ and $\mathbf{Q} = \mathbf{P}_{[N]} = \mathbf{I}$, we obtain

$$\|\mathbf{P}_{0I}\mathbf{K}\|_{\mathrm{HS}(\mathcal{H})}^2 = \|\mathbf{P}_{0I}\mathbf{KI}\|_{\mathrm{HS}(\mathcal{H})}^2 = \langle \mathbf{P}_{0I}\mathbf{K} \,|\, \mathbf{IK}\rangle_{\mathrm{F}} = \langle \mathbf{K} - \hat{\mathbf{K}}(I) \,|\, \mathbf{K}\rangle_{\mathrm{F}} \tag{3.15}$$

Similarly, we have (with $\mathrm{Re}(z)$ the real part of $z \in \mathbb{C}$)

$$\begin{aligned}\|\mathbf{K} - \mathbf{P}_I\mathbf{KP}_I\|_{\mathrm{HS}(\mathcal{H})}^2 &= \|\mathbf{K}\|_{\mathrm{HS}(\mathcal{H})}^2 - 2\mathrm{Re}(\langle \mathbf{K} \,|\, \mathbf{P}_I\mathbf{KP}_I\rangle_{\mathrm{HS}(\mathcal{H})}) + \|\mathbf{P}_I\mathbf{KP}_I\|_{\mathrm{HS}(\mathcal{H})}^2 \\ &= \|\mathbf{K}\|_{\mathrm{F}}^2 - \|\mathbf{P}_I\mathbf{KP}_I\|_{\mathrm{HS}(\mathcal{H})}^2 = \|\mathbf{K}\|_{\mathrm{F}}^2 - \|\hat{\mathbf{K}}(I)\|_{\mathrm{F}}^2,\end{aligned} \tag{3.16}$$

where we have applied Lemma 3.2 with $\mathbf{P} = \mathbf{Q} = \mathbf{P}_I$, and we have used the fact that for any orthogonal projection $\mathbf{P}$ on $\mathcal{H}$,

$$\langle \mathbf{K} \,|\, \mathbf{PKP}\rangle_{\mathrm{HS}(\mathcal{H})} = \|\mathbf{PKP}\|_{\mathrm{HS}(\mathcal{H})}^2. \tag{3.17}$$

Remark 3.5. By interpreting Nyström approximations of $\mathbf{K}$ as HS operators on $\mathcal{H}$, we can relate the low-rank approximation of PSD matrices to the approximation of potentials in squared-kernel RKHSs. This allows for the definition of energy-based surrogate error maps for column sampling; see Chapters 4 and 5.

Case 4: $\mathcal{H} \to \mathcal{E}$. Any matrix $\mathbf{M} \in \mathbb{C}^{N\times N}$ can be regarded as an operator from $\mathcal{H}$ to $\mathcal{E}$, and in this case, we in particular have

$$\|\mathbf{K} - \mathbf{P}_I\mathbf{KP}_I\|_{\mathrm{HS}(\mathcal{H},\mathcal{E})} = \mathrm{trace}(\mathbf{K}^3) + \mathrm{trace}((\hat{\mathbf{K}}(I))^2(\hat{\mathbf{K}}(I) - 2\mathbf{K}));$$

further details on this norm can be found in the work of Gauthier (2024). Due to its high computational complexity, we will not consider this case in the manuscript.

3.2.2 Error maps and Hilbert–Schmidt norms

Following Section 3.2.1, we define the following maps, relating to the trace, Frobenius and spectral norm errors

(C.1) $\mathcal{C}_{\mathrm{tr}}(I) = \|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{tr}}$, the trace error map;

(C.2) $\mathcal{C}_{\mathrm{F}}(I) = \|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{F}}^2$, the Frobenius error map;

(C.3) $\mathcal{C}_{\mathrm{sp}}(I) = \|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{sp}}^2$, the spectral error map.

For $\mathcal{C}_{\mathrm{tr}}$, note that the norm is not squared, as in (3.11). We also introduce two more error maps based on (3.15) and (3.16):

(C.4) $\mathcal{C}_{\mathrm{P}}(I) = \langle \mathbf{K} - \hat{\mathbf{K}}(I) \,|\, \mathbf{K} \rangle_{\mathrm{F}}$, the projection error map;

(C.5) $\mathcal{C}_{\mathrm{PP}}(I) = \|\mathbf{K}\|_{\mathrm{F}}^2 - \|\hat{\mathbf{K}}(I)\|_{\mathrm{F}}^2$, the double-projection error map.

Lemma 3.3. For all $I \subseteq [N]$, the following inequalities hold:

$$\mathcal{C}_{\mathrm{sp}}(I) \leqslant \mathcal{C}_{\mathrm{F}}(I) \leqslant \mathcal{C}_{\mathrm{P}}(I) \leqslant \mathcal{C}_{\mathrm{PP}}(I).$$

Proof. The inequality $\mathcal{C}_{\mathrm{sp}}(I) \leqslant \mathcal{C}_{\mathrm{F}}(I)$ follows from the relation between the Frobenius and spectral norms. From Lemma 3.2, we have

$$\begin{aligned} \mathcal{C}_{\mathrm{F}}(I) &= \|\mathbf{K}\|_{\mathrm{F}}^2 + \|\hat{\mathbf{K}}(I)\|_{\mathrm{F}}^2 - 2\mathrm{Re}\Big(\langle \hat{\mathbf{K}}(I) \,|\, \mathbf{K} \rangle_{\mathrm{F}}\Big) \\ &= \|\mathbf{K}\|_{\mathrm{F}}^2 + \|\mathbf{P}_I \mathbf{K} \mathbf{P}_I\|_{\mathrm{HS}(\mathcal{H})}^2 - 2\|\mathbf{P}_I \mathbf{K}\|_{\mathrm{HS}(\mathcal{H})}^2. \end{aligned} \tag{3.18}$$

The matrix $\mathbf{P}_{0I}$ corresponds to the orthogonal projection from $\mathcal{H}$ onto the orthogonal complement of $\mathcal{H}_I$ in $\mathcal{H}$, and so

$$\|\mathbf{P}_I \mathbf{K}\|_{\mathrm{HS}(\mathcal{H})}^2 = \|\mathbf{P}_I \mathbf{K} \mathbf{P}_I\|_{\mathrm{HS}(\mathcal{H})}^2 + \|\mathbf{P}_I \mathbf{K} \mathbf{P}_{0I}\|_{\mathrm{HS}(\mathcal{H})}^2 \geqslant \|\mathbf{P}_I \mathbf{K} \mathbf{P}_I\|_{\mathrm{HS}(\mathcal{H})}^2. \tag{3.19}$$

Combining (3.18) and (3.19), we obtain

$$\mathcal{C}_{\mathrm{F}}(I) \leqslant \|\mathbf{K}\|_{\mathrm{F}}^2 - \|\mathbf{P}_I \mathbf{K}\|_{\mathrm{HS}(\mathcal{H})}^2 = \mathcal{C}_{\mathrm{P}}(I) \leqslant \|\mathbf{K}\|_{\mathrm{F}}^2 - \|\mathbf{P}_I \mathbf{K} \mathbf{P}_I\|_{\mathrm{HS}(\mathcal{H})}^2 = \mathcal{C}_{\mathrm{PP}}(I),$$

completing the proof. □

In Chapter 4, further properties of these projection-based error maps will be explored; indeed, we will see that they are closely related to the proposed energy-based optimisation framework.

Chapter 4

Sequential column sampling strategies

In this chapter, we propose a class of sequential sampling strategies for the Nyström method which leverage the properties of a differentiable pseudoconvex relaxation of the column-sampling problem (CSP), in which samples of columns are characterised through the non-zero entries of *selection vectors*.

The chapter is organised as follows. In Section 4.1, we describe the overall framework surrounding the considered relaxation of the CSP, and introduce an energy-based error map R on the selection-vector space that is differentiable and pseudoconvex. In Section 4.2, we present a class of sequential column-sampling strategies which utilise the gradient of the error map R; stochastic variants of these strategies are discussed in Section 4.3. Section 4.4 is devoted to numerical experiments, and an additional figure for the experiment in Section 4.4.2 is provided in the appendix of this chapter.

4.1 Relaxing the column-sampling problem

In this section, we describe and discuss a number of relaxations to the CSP. Firstly, we introduce the notion of selection vectors, and show that this alone leads to a convex, but non-differentiable, relaxation. Following this, we present an error map defined on the selection-vector space that is both convex and differentiable. Finally, we introduce an invariance with respect to rescaling of the input vectors, leading to a pseudoconvex differentiable error map, on which the proposed sampling strategies in Section 4.2 are based.

As in Section 3.2, we let $\mathbf{K} \in \mathbb{C}^{N\times N}$ be a PSD matrix, and we identify the RKHS $\mathcal{H}$ as the vector subspace $\mathcal{H} = \operatorname{span}\{\mathbf{K}\} \subseteq \mathbb{C}^N$ equipped with the inner product $\langle \cdot \,|\, \cdot \rangle_{\mathcal{H}}$ given by

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

4.1.1 First relaxation: selection vectors

For $\boldsymbol{\upsilon} = (\upsilon_i)_{i\in[N]} \in \mathbb{R}^N$, we set $I_{\boldsymbol{\upsilon}} = \{i \in [N] \,|\, \upsilon_i \neq 0\}$, and we refer to $I_{\boldsymbol{\upsilon}}$ as the *support* of $\boldsymbol{\upsilon}$. Through its support, a *selection vector* $\boldsymbol{\upsilon}$ characterises a subset of columns of $\mathbf{K}$; following Section 3.2, we introduce the simplified notations

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

We then define the following error maps on $\mathbb{R}^N$, which are analogous to the maps $\mathcal{C}_{\mathrm{tr}}$, $\mathcal{C}_{\mathrm{F}}$ and $\mathcal{C}_{\mathrm{sp}}$ in Chapter 3:

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

Theorem 4.1. 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 $\boldsymbol{\upsilon}, \boldsymbol{\eta} \in \mathbb{R}^N_{\geqslant 0}$, 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 $\{-\infty, 0\}$.

Theorem 4.1 illustrates that the error maps on the selection-vector space induced by the trace, Frobenius and spectral norms are akin to convex piecewise-constant functions on $\mathbb{R}^N_{\geqslant 0}$; see Figure 4.1 for an illustration. The selection-vector formulation can hence be regarded as a *nondifferentiable convex relaxation* of the CSP. Introducing $|\boldsymbol{\upsilon}| = (|\upsilon_i|)_{i\in[N]} \in \mathbb{R}^N_{\geqslant 0}$, we observe that $C_{\mathrm{X}}(\boldsymbol{\upsilon}) = C_{\mathrm{X}}(|\boldsymbol{\upsilon}|)$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$.

The following Lemma 4.1 is a supporting result for the proof of Theorem 4.1.

Lemma 4.1. For $J \subseteq I \subseteq [N]$, we have

$$\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{X}} \leqslant \|\mathbf{K} - \hat{\mathbf{K}}(J)\|_{\mathrm{X}}, \quad \mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}.$$

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

Trace norm. From (3.11), and noticing that $\langle \mathbf{P}_e\mathbf{K} \,|\, \mathbf{P}_{0I}\mathbf{K}\rangle_{\mathrm{HS}(\mathcal{E},\mathcal{H})} = 0$, we have

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

Frobenius norm. Since $\mathcal{H}_{0I}$ and $\mathcal{H}_e$ are orthogonal in $\mathcal{H}$, the matrices $\mathbf{P}_{0I}\mathbf{K}\mathbf{P}_{0I}$, $\mathbf{P}_e\mathbf{K}\mathbf{P}_e$, $\mathbf{P}_{0I}\mathbf{K}\mathbf{P}_e$ and $\mathbf{P}_e\mathbf{K}\mathbf{P}_{0I}$ are orthogonal in $\mathrm{HS}(\mathcal{H})$. From (3.12) and using Lemma 3.2, we obtain

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

Spectral norm. We first observe that if $\mathbf{P} \in \mathbb{C}^{N\times N}$ is an orthogonal projection on $\mathcal{H}$, then the PSD operator on $\mathcal{E}$ related to $\mathbf{PK}$ and the PSD operator on $\mathcal{H}$ related to $\mathbf{PKP}$ have the same strictly-positive eigenvalues. Indeed, if $\mathbf{PKv} = \lambda\mathbf{v}$, with $\mathbf{v} \in \mathcal{E}$, $\mathbf{v} \neq 0$, and $\lambda > 0$, then

$$\lambda\mathbf{P}\mathbf{v} = \mathbf{P}(\lambda\mathbf{v}) = \mathbf{PPKv} = \mathbf{PKv} = \lambda\mathbf{v},$$

and so $\lambda(\mathbf{Pv} - \mathbf{v}) = 0$; as $\lambda > 0$, we obtain $\mathbf{v} = \mathbf{Pv} \in \mathcal{H}$ and $\mathbf{PKPv} = \lambda\mathbf{v}$. Conversely, if $\mathbf{PKP}\boldsymbol{h} = \lambda\boldsymbol{h}$, with $\boldsymbol{h} \in \mathcal{H}$, $\boldsymbol{h} \neq 0$ and $\lambda > 0$, then

$$\lambda\mathbf{P}\boldsymbol{h} = \mathbf{P}(\lambda\boldsymbol{h}) = \mathbf{PPKP}\boldsymbol{h} = \mathbf{PKP}\boldsymbol{h} = \lambda\boldsymbol{h},$$

and so $\lambda(\mathbf{P}\boldsymbol{h} - \boldsymbol{h}) = 0$; as $\lambda > 0$, we have $\mathbf{P}\boldsymbol{h} = \boldsymbol{h}$ and $\mathbf{PK}\boldsymbol{h} = \lambda\boldsymbol{h}$.

For the spectral norm error, observing that $\mathcal{H}_{0I} \subseteq \mathcal{H}_{0J}$, we get

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

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

completing the proof. □

We now prove Theorem 4.1.

Proof of Theorem 4.1. For $\boldsymbol{\xi} = \boldsymbol{\upsilon} + \rho(\boldsymbol{\eta} - \boldsymbol{\upsilon})$, $\boldsymbol{\upsilon}, \boldsymbol{\eta} \in \mathbb{R}^N_{\geqslant 0}$, $\rho \in (0,1)$, we have $I_{\boldsymbol{\xi}} = I_{\boldsymbol{\upsilon}} \cup I_{\boldsymbol{\eta}}$, 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 4.1, 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. □

4.1.2 Second relaxation: quadrature approximation

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 $\mathrm{HS}(\mathcal{H})$. Let $\boldsymbol{\omega} \in \mathbb{R}^N$ be another selection vector, and set $\mathbf{W} = \mathrm{diag}(\boldsymbol{\omega})$. From (3.14), we have

$$
\langle \mathbf{K}\mathbf{W} \,|\, \mathbf{K}\mathbf{V}\rangle_{\mathrm{HS}(\mathcal{H})} = \mathrm{trace}(\mathbf{K}\mathbf{W}\mathbf{K}\mathbf{K}^{\dagger}\mathbf{K}\mathbf{V}) = \mathrm{trace}(\mathbf{K}\mathbf{W}\mathbf{K}\mathbf{V}) = \boldsymbol{\omega}^{*}\mathbf{S}\boldsymbol{\upsilon}, \tag{4.1}
$$

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 $\mathbf{K}$; we remark that the matrix $\mathbf{S}$ is real, symmetric and PSD. Introducing $\mathbb{1} = (1)_{i \in [N]} \in \mathbb{R}^N$, we in particular have $\mathrm{diag}(\mathbb{1}) = \mathbf{I}$ 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

$$
\begin{aligned}
D(\boldsymbol{\upsilon}) &= \|\mathbf{K} - \mathbf{K}\mathbf{V}\|^2_{\mathrm{HS}(\mathcal{H})} = (\mathbb{1} - \boldsymbol{\upsilon})^{*}\mathbf{S}(\mathbb{1} - \boldsymbol{\upsilon}) \\
&= \|\mathbf{K}\|^2_{\mathrm{F}} + \boldsymbol{\upsilon}^{*}\mathbf{S}\boldsymbol{\upsilon} - 2\mathfrak{g}^{*}\boldsymbol{\upsilon}, \quad \boldsymbol{\upsilon} \in \mathbb{R}^N,
\end{aligned} \tag{4.2}
$$

with $\mathfrak{g} = \mathbf{S}\mathbb{1} \in \mathbb{R}^N_{\geqslant 0}$. We refer to $\mathfrak{g}$ as the *target potential* (see Remark 4.1), since this corresponds to sampling every column of $\mathbf{K}$, the target matrix, and assigning them all equal weight.

Lemma 4.2. The error map D is convex on $\mathbb{R}^N$, and the gradient of D at $\boldsymbol{\upsilon}$ is given by $\nabla D(\boldsymbol{\upsilon}) = 2(\mathbf{S}\boldsymbol{\upsilon} - \mathfrak{g})$.

Proof. For $\boldsymbol{\upsilon} \in \mathbb{R}^N$, and using standard differentiation rules, from (4.2) we obtain

$$\nabla D(\boldsymbol{\upsilon}) = \nabla \|\mathbf{K}\|_{\mathrm{F}}^2 + \nabla \boldsymbol{\upsilon}^* \mathbf{S} \boldsymbol{\upsilon} - \nabla 2\mathfrak{g}^* \boldsymbol{\upsilon} = 0 + 2\mathbf{S}\boldsymbol{\upsilon} - 2\mathfrak{g} = 2(\mathbf{S}\boldsymbol{\upsilon} - \mathfrak{g}).$$

The map D is convex on $\mathbb{R}^N$ if and only if for all $\boldsymbol{\upsilon}, \boldsymbol{\eta} \in \mathbb{R}^N$,

$$D(\boldsymbol{\upsilon}) - D(\boldsymbol{\eta}) - \nabla D(\boldsymbol{\eta})^*(\boldsymbol{\upsilon} - \boldsymbol{\eta}) \geqslant 0.$$

Let $\boldsymbol{\upsilon}, \boldsymbol{\eta} \in \mathbb{R}^N$. We have

$$\begin{aligned} D(\boldsymbol{\upsilon}) - D(\boldsymbol{\eta}) - \nabla D(\boldsymbol{\eta})^*(\boldsymbol{\upsilon} - \boldsymbol{\eta}) &= \boldsymbol{\upsilon}^* \mathbf{S} \boldsymbol{\upsilon} - 2\mathfrak{g}^* \boldsymbol{\upsilon} - \boldsymbol{\eta}^* \mathbf{S} \boldsymbol{\eta} + 2\mathfrak{g}^* \boldsymbol{\eta} \\ &\quad - 2(\mathbf{S}\boldsymbol{\eta} - \mathfrak{g})^*(\boldsymbol{\upsilon} - \boldsymbol{\eta}) \\ &= \boldsymbol{\upsilon}^* \mathbf{S} \boldsymbol{\upsilon} - 2\boldsymbol{\eta}^* \mathbf{S} \boldsymbol{\upsilon} + \boldsymbol{\eta}^* \mathbf{S} \boldsymbol{\eta} \\ &= (\boldsymbol{\upsilon} - \boldsymbol{\eta})^* \mathbf{S} (\boldsymbol{\upsilon} - \boldsymbol{\eta}) \geqslant 0, \end{aligned}$$

where the inequality follows from the fact that $\mathbf{S}$ is PSD. □

Remark 4.1. The PSD matrix $\mathbf{S}$ defines a RKHS that can be identified with the vector space $\mathcal{G} = \operatorname{span}\{\mathbf{S}\} \subseteq \mathbb{C}^N$ endowed with the inner product $\langle \boldsymbol{g} \,|\, \boldsymbol{j} \rangle_{\mathcal{G}} = \boldsymbol{g}^* \mathbf{S}^\dagger \boldsymbol{j}$ for $\boldsymbol{g}, \boldsymbol{j} \in \mathcal{G}$. In view of (4.1), we have

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

We refer to $\mathbf{S}\boldsymbol{\upsilon}$ as the *potential* of $\boldsymbol{\upsilon}$ in $\mathcal{G}$, and to $\|\mathbf{S}\boldsymbol{\upsilon}\|_{\mathcal{G}}^2 = \|\mathbf{K}\mathbf{V}\|_{\mathrm{HS}(\mathcal{H})}^2 = \boldsymbol{\upsilon}^* \mathbf{S} \boldsymbol{\upsilon}$ as the *energy* of $\boldsymbol{\upsilon}$ with respect to $\mathbf{S}$. In fact, these correspond to g_ν and $\|g_\nu\|_{\mathcal{G}}^2$ in the energy setting of Section 3.1.3 when ν is taken to be the discrete measure $\sum_{i=1}^N \upsilon_i \delta_i$. The energy-based error map D on $\mathbb{R}^N$ then corresponds to the square of the MMD $(\mu, \nu) \mapsto \|L_\mu - L_\nu\|_{\mathrm{HS}(\mathcal{H})}$, $\mu, \nu \in \mathcal{T}(K)$ (see Section 3.1.3.2). ◁

4.1.3 Invariance under rescaling

For $\boldsymbol{\upsilon} \in \mathbb{R}^N$ and $c > 0$, we have $I_{\boldsymbol{\upsilon}} = I_{c\boldsymbol{\upsilon}}$; 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 (4.2), we introduce the error map

$$R(\boldsymbol{\upsilon}) = \min_{c \geqslant 0} D(c\boldsymbol{\upsilon}) = \begin{cases} \|\mathbf{K}\|_{\mathrm{F}}^2 - (\mathfrak{g}^* \boldsymbol{\upsilon})^2 / (\boldsymbol{\upsilon}^* \mathbf{S} \boldsymbol{\upsilon}) \text{ if } \mathfrak{g}^* \boldsymbol{\upsilon} > 0, \\ \|\mathbf{K}\|_{\mathrm{F}}^2 \text{ otherwise,} \end{cases} \tag{4.3}$$

and we set $\mathscr{D} = \{\boldsymbol{v} \in \mathbb{R}^N \,|\, \mathfrak{g}^*\boldsymbol{v} > 0\}$.

The appearance of the error maps D, R and C_{F} in two dimensions is illustrated in Figure 4.1.

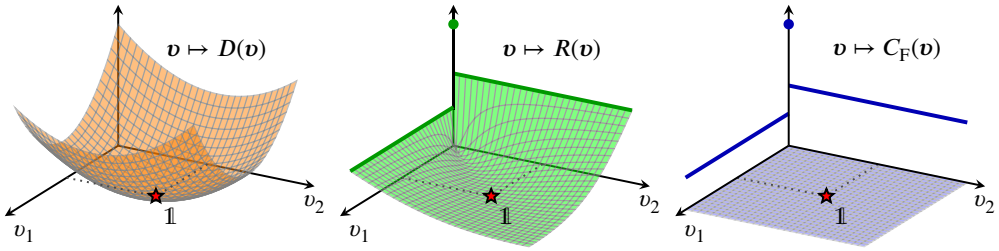

Figure 4.1: Schematic representation of the error maps D, R and C_{F} on $\mathbb{R}^N_{\geqslant 0}$; the red star represents the selection vector $\mathbb{1} \in \mathbb{R}^N$. The presented graphs correspond to a 2×2 PSD matrix $\mathbf{K}$ such that $\mathbf{K}_{1,1} = 1.225$, $\mathbf{K}_{2,2} = 0.894$ and $\mathbf{K}_{2,1} = \mathbf{K}_{1,2} = 0.316$. 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. The scaling invariance of R is further illustrated by the fact that the surface is flat (and minimal) along the line passing through the origin and the red star at $\boldsymbol{v} = \mathbb{1}$.

From the Cauchy-Schwarz 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

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

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

$$\begin{aligned} \Theta(\boldsymbol{v}; \boldsymbol{\eta}) &= \lim_{\rho \to 0^+} \frac{1}{\rho}\Big[R\Big(\boldsymbol{v} + \rho(\boldsymbol{\eta} - \boldsymbol{v})\Big) - R(\boldsymbol{v})\Big] \\ &= \begin{cases} -\infty \text{ if } \boldsymbol{v} \in \mathscr{Z} \text{ and } \boldsymbol{\eta} \in \mathscr{D}, \\ 2c_{\boldsymbol{v}}(\boldsymbol{\eta} - \boldsymbol{v})^*(c_{\boldsymbol{v}}\mathbf{S}\boldsymbol{v} - \mathfrak{g}) \text{ otherwise,} \end{cases} \end{aligned} \tag{4.4}$$

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 all $\boldsymbol{v} \in \mathscr{D}$, $\boldsymbol{v}^*(c_{\boldsymbol{v}}\mathbf{S}\boldsymbol{v} - \mathfrak{g}) = 0$.

Theorem 4.2. The map R is quasiconvex on $\mathbb{R}^N$, and pseudoconvex on the convex cone $\mathscr{D}$.

Proof. We first show the quasiconvexity of R on $\mathbb{R}^N$. For $\boldsymbol{\xi} = \boldsymbol{v} + \rho(\boldsymbol{\eta} - \boldsymbol{v})$, $\boldsymbol{v}, \boldsymbol{\eta} \in \mathbb{R}^N$, $\rho \in [0, 1]$, there always exists $c \geqslant 0$ and $\rho' \in [0, 1]$ such that $c\boldsymbol{\xi} = (1 - \rho')c_{\boldsymbol{v}}\boldsymbol{v} + \rho' c_{\boldsymbol{\eta}}\boldsymbol{\eta}$; indeed:

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

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

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

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,

$$(\boldsymbol{\upsilon}^* \mathbf{S}\mathbb{1})(\boldsymbol{\eta}^* \mathbf{S}\boldsymbol{\upsilon}) \geqslant (\boldsymbol{\upsilon}^* \mathbf{S}\boldsymbol{\upsilon})(\boldsymbol{\eta}^* \mathbf{S}\mathbb{1}). \tag{4.5}$$

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 (4.5), $\boldsymbol{\eta}^* \mathbf{S}\boldsymbol{\upsilon} > 0$. The matrix $\mathbf{S}$ being PSD, the Cauchy-Schwarz inequality gives $(\boldsymbol{\eta}^* \mathbf{S}\boldsymbol{\upsilon})^2 \leqslant (\boldsymbol{\upsilon}^* \mathbf{S}\boldsymbol{\upsilon})(\boldsymbol{\eta}^* \mathbf{S}\boldsymbol{\eta})$; combining this inequality with (4.5), 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 $\mathscr{D}$. □

For $\boldsymbol{\upsilon}^\star = c\mathbb{1} + \boldsymbol{\epsilon}$, with $c > 0$ and $\boldsymbol{\epsilon} \in \mathbb{R}^N$ such that $\mathbf{S}\boldsymbol{\epsilon} = 0$, we have $R(\boldsymbol{\upsilon}^\star) = 0$, and R is thus minimum at $\boldsymbol{\upsilon}^\star$. 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 $\mathscr{D}$ (see e.g. Lee et al. (2016)). Lemma 4.3 provides an analytical expression for the optimal step size and for the improvement induced by a descent with optimal step size in the case of interest for Section 4.2.

Lemma 4.3. For $\boldsymbol{\upsilon} \in \mathscr{D}$ and $\boldsymbol{\eta} \in \mathbb{R}^N$ such that $\Theta(\boldsymbol{\upsilon};\boldsymbol{\eta}) < 0$ and $\Theta(\boldsymbol{\eta};\boldsymbol{\upsilon}) \leqslant 0$, the function $\rho \mapsto R\Big(\boldsymbol{\upsilon} + \rho(\boldsymbol{\eta} - \boldsymbol{\upsilon})\Big)$, $\rho \in [0,1]$, is minimum at $\rho = r \in (0,1]$, with

$$r = \frac{T_1}{T_1 + T_2}, \tag{4.6}$$

where $T_1 = (\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon})(\mathfrak{g}^*\boldsymbol{\eta}) - (\mathfrak{g}^*\boldsymbol{\upsilon})(\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\eta})$ and $T_2 = (\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta})(\mathfrak{g}^*\boldsymbol{\upsilon}) - (\mathfrak{g}^*\boldsymbol{\eta})(\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\eta})$; introducing $\mathcal{I}(\boldsymbol{\upsilon};\boldsymbol{\eta}) = R(\boldsymbol{\upsilon}) - R\Big(\boldsymbol{\upsilon} + r(\boldsymbol{\eta} - \boldsymbol{\upsilon})\Big) \geqslant 0$, we have

$$\mathcal{I}(\boldsymbol{\upsilon};\boldsymbol{\eta}) = \frac{\Big(\boldsymbol{\eta}^*(c_{\boldsymbol{\upsilon}}\mathbf{S}\boldsymbol{\upsilon} - \mathfrak{g})\Big)^2}{\Big((\boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta}) - (\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\eta})^2/(\boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon})\Big)}. \tag{4.7}$$

Proof. We set $a = \mathfrak{g}^*\boldsymbol{\upsilon} > 0$, $b = \mathfrak{g}^*\boldsymbol{\eta}$, $c = \boldsymbol{\upsilon}^*\mathbf{S}\boldsymbol{\upsilon} > 0$, $d = \boldsymbol{\eta}^*\mathbf{S}\boldsymbol{\eta}$, 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

$$\varphi(x) = \mathfrak{g}^*\boldsymbol{\xi}(x) = x(b-a) + a$$

and

$$\psi(x) = \boldsymbol{\xi}(x)^*\mathbf{S}\boldsymbol{\xi}(x) = x^2(c + d - 2e) + 2x(e - c) + c.$$

The condition $\Theta(\boldsymbol{\upsilon};\boldsymbol{\eta}) < 0$ ensures that the degree-2 polynomial ψ is strictly positive; indeed, ψ is nonnegative and admits a real root if and only if $e^2 = cd$, that is, from the Cauchy-Schwarz 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 $f(x) = -\varphi^2(x)/\psi(x)$ for all $x \in \mathbb{R}$, so that if $\boldsymbol{\xi}(x) \in \mathscr{D}$, then we have $f(x) = R(\boldsymbol{\xi}(x)) - \|\mathbf{K}\|_{\mathrm{F}}^2$. The derivative of f is given by

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

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 $\mathfrak{g}^*\boldsymbol{\xi}(x_1) = 0$, and so $R(\boldsymbol{\xi}(x_1)) = \|\mathbf{K}\|_{\mathrm{F}}^2 > R(\boldsymbol{\upsilon})$; we can therefore conclude that $r \neq x_1$. Cancelling the linear function $x \mapsto x\Big((bc - ae) + (ad - be)\Big) - (bc - ae)$, we obtain $f'(x_2) = 0$ with

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

we therefore necessarily have $r = x_2$, and

$$\mathcal{I}(\boldsymbol{\upsilon};\boldsymbol{\eta}) = f(0) - f(x_2) = \frac{(bc-ae)^2}{c(cd-e^2)},$$

as required. □

4.1.4 Additional error maps and further properties

In Section 3.2.2, we introduced two projection-based error maps $\mathcal{C}_{\mathrm{P}}$ and $\mathcal{C}_{\mathrm{PP}}$. We now define their corresponding error maps in the selection-vector setting:

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

The maps C_{P} and C_{PP} are of the same type as the maps C_{X}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}$, as illustrated by Proposition 4.1.

Proposition 4.1. The maps C_{P} and C_{PP} are convex on the convex cone $\mathbb{R}^N_{\geqslant 0}$, and their directional derivatives take values in the discrete set $\{-\infty, 0\}$.

Proof. We follow the proof of Theorem 4.1, and show that if $J \subseteq I \subseteq [N]$, then

$$\|\mathbf{K} - \mathbf{P}_I\mathbf{K}\|_{\mathrm{HS}(\mathcal{H})} \leqslant \|\mathbf{K} - \mathbf{P}_J\mathbf{K}\|_{\mathrm{HS}(\mathcal{H})}$$

and

$$\|\mathbf{K} - \mathbf{P}_I\mathbf{K}\mathbf{P}_I\|_{\mathrm{HS}(\mathcal{H})} \leqslant \|\mathbf{K} - \mathbf{P}_J\mathbf{K}\mathbf{P}_J\|_{\mathrm{HS}(\mathcal{H})}.$$

Using the same notations as in the proof of Lemma 4.1, and noticing that $\mathcal{H}_{0I}$ and $\mathcal{H}_e$ are orthogonal in $\mathcal{H}$, we have

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

as required. Next, by (3.17),

$$\|\mathbf{K} - \mathbf{P}\mathbf{K}\mathbf{P}\|_{\mathrm{HS}(\mathcal{H})}^2 = \|\mathbf{K}\|_{\mathrm{HS}(\mathcal{H})}^2 - \|\mathbf{P}\mathbf{K}\mathbf{P}\|_{\mathrm{HS}(\mathcal{H})}^2. \tag{4.8}$$

Observing that the matrices $\mathbf{P}_J\mathbf{K}\mathbf{P}_J$, $\mathbf{P}_e\mathbf{K}\mathbf{P}_e$, $\mathbf{P}_J\mathbf{K}\mathbf{P}_e$ and $\mathbf{P}_e\mathbf{K}\mathbf{P}_J$ are orthogonal in $\mathrm{HS}(\mathcal{H})$, we obtain

$$\begin{aligned}\|\mathbf{P}_I\mathbf{K}\mathbf{P}_I\|_{\mathrm{HS}(\mathcal{H})}^2 &= \|\mathbf{P}_J\mathbf{K}\mathbf{P}_J\|_{\mathrm{HS}(\mathcal{H})}^2 + \|\mathbf{P}_e\mathbf{K}\mathbf{P}_J\|_{\mathrm{HS}(\mathcal{H})}^2 \\ &\quad + \|\mathbf{P}_J\mathbf{K}\mathbf{P}_e\|_{\mathrm{HS}(\mathcal{H})}^2 + \|\mathbf{P}_e\mathbf{K}\mathbf{P}_e\|_{\mathrm{HS}(\mathcal{H})}^2 \\ &\geqslant \|\mathbf{P}_J\mathbf{K}\mathbf{P}_J\|_{\mathrm{HS}(\mathcal{H})}^2,\end{aligned}$$

giving, in combination with (4.8), the expected inequality. □

The following Lemma 4.4 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 notice that

$$C_{\mathrm{X}}(\mathbb{1}) = R(\mathbb{1}) = 0, \quad \mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}, \mathrm{P}, \mathrm{PP}\},$$

and

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

Lemma 4.4. 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, $C_{\mathrm{PP}}(\mathbf{e}_i) = R(\mathbf{e}_i)$ for all $i \in [N]$.

Proof. The chain of inequalities $C_{\mathrm{sp}}(\boldsymbol{\upsilon}) \leqslant C_{\mathrm{F}}(\boldsymbol{\upsilon}) \leqslant C_{\mathrm{P}}(\boldsymbol{\upsilon}) \leqslant C_{\mathrm{PP}}(\boldsymbol{\upsilon})$ follows directly from Lemma 3.3. For the remainder, we begin by observing that $\mathbf{KV}\boldsymbol{h} = \mathbf{P}_{\boldsymbol{\upsilon}}\mathbf{KV}\mathbf{P}_{\boldsymbol{\upsilon}}\boldsymbol{h}$ for all $\boldsymbol{h} \in \mathcal{H}$ (indeed, we have $\mathrm{span}\{\mathbf{KV}\} \subseteq \mathcal{H}_{\boldsymbol{\upsilon}}$, and $\mathbf{e}_i^*\mathbf{P}_{\boldsymbol{\upsilon}}\boldsymbol{h} = \mathbf{e}_i^*\boldsymbol{h}$ for all $i \in I_{\boldsymbol{\upsilon}}$), and so $\langle \mathbf{K} - \mathbf{P}_{\boldsymbol{\upsilon}}\mathbf{K}\mathbf{P}_{\boldsymbol{\upsilon}} \,|\, \mathbf{P}_{\boldsymbol{\upsilon}}\mathbf{K}\mathbf{P}_{\boldsymbol{\upsilon}} - \mathbf{KV} \rangle_{\mathrm{HS}(\mathcal{H})} = 0$. We hence obtain

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

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

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

In view of the above developments, we propose to use the error map R as a differentiable surrogate for the characterisation of samples of columns for Nyström approximation through the supports of selection vectors. In the forthcoming Section 4.2, we describe a class of sequential sampling strategies driven by the gradient of R.

4.2 Gradient-based sequential sampling

From now on, we assume that the diagonal entries of $\mathbf{K}$ are strictly positive, so that $\mathbb{R}_{\geqslant 0}^N \setminus \{0\} \subset \mathscr{D}$ (this assumption is nonrestrictive: if a diagonal entry of $\mathbf{K}$ is

zero, then by Cauchy-Schwarz, the corresponding row and column of $\mathbf{K}$ are also zero). 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 descent directions (specifically, the directions defined by the extreme points of $\mathcal{A}_{\mathbf{f}}$). Many variants could be considered, see for instance Remarks 4.2, 4.3 and 4.4; stochastic variants are discussed in Section 4.3. 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

$$b \in \arg\min_{i\in[N]} R(\boldsymbol{\xi}_i) = \arg\max_{i\in[N]} \mathfrak{g}_i^2/\mathbf{S}_{i,\,i}, \tag{4.9}$$

with $\mathfrak{g}_i = \mathbf{e}_i^*\mathfrak{g}$ the i-th entry of $\mathfrak{g} = \mathbf{S}\mathbb{1}$, and the selection vector at step $q \in \mathbb{N}$ is denoted by $\boldsymbol{\upsilon}^{(q)} \in \mathcal{A}_{\mathbf{f}}$. An iteration of our sampling procedure consists of selecting a descent direction $\boldsymbol{\xi}_u - \boldsymbol{\upsilon}^{(q)}$, with $u \in [N]$ such that $\Theta(\boldsymbol{\upsilon}^{(q)}; \boldsymbol{\xi}_u) < 0$, and of next performing a descent with the optimal step size r given by (4.6). As descent direction, we consider the Frank-Wolfe (FW) direction $\boldsymbol{\xi}_u - \boldsymbol{\upsilon}^{(q)}$, with

$$u \in \arg\min_{i\in[N]} \Theta(\boldsymbol{\upsilon}^{(q)}; \boldsymbol{\xi}_i) = \arg\min_{i\in[N]} [\nabla R(\boldsymbol{\upsilon}^{(q)})]_i/f_i. \tag{4.10}$$

The initialisation of the procedure via (4.9) ensures that if $\Theta(\boldsymbol{\upsilon}^{(q)}; \boldsymbol{\xi}_i) < 0$, $i \in [N]$, then $\Theta(\boldsymbol{\xi}_i; \boldsymbol{\upsilon}^{(q)}) < 0$, so that the descents necessarily occur within the framework of Lemma 4.3.

A pseudocode of the procedure is given in Algorithm 4.1. The algorithm produces a sequence $\boldsymbol{\upsilon}^{(1)}, \boldsymbol{\upsilon}^{(2)}, \ldots$ 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 $m_q \leqslant \min(q, N)$, so that early stopping ensures sparsity of the resulting selection vector (see also Remark 4.3). The algorithm stops if $\boldsymbol{\upsilon}^{(q)}$ minimises R over $\mathcal{A}_{\mathbf{f}}$, or 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). We observe that $\boldsymbol{\upsilon}^\star = \varkappa\mathbb{1}/(\mathbf{f}^*\mathbb{1}) \in \mathcal{A}_{\mathbf{f}}$ verifies $R(\boldsymbol{\upsilon}^\star) = 0$.

Algorithm 4.1: 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 4.3);
Initialisation: compute $b \in [N]$ using (4.9); set $q = 1$, $\boldsymbol{\upsilon}^{(1)} = \boldsymbol{\xi}_b$ and $I_{\boldsymbol{\upsilon}^{(1)}} = \{b\}$;
while $q < Q$ *and* $R(\boldsymbol{\upsilon}^{(q)}) > 0$ **do**
  compute $u \in [N]$ using (4.10);
  compute the optimal step size r from (4.6) with $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}^{(q)}$ and $\boldsymbol{\eta} = \boldsymbol{\xi}_u$;
  set $\boldsymbol{\upsilon}^{(q+1)} = (1-r)\boldsymbol{\upsilon}^{(q)} + r\boldsymbol{\xi}_u$ and $I_{\boldsymbol{\upsilon}^{(q+1)}} = I_{\boldsymbol{\upsilon}^{(q)}} \cup \{u\}$; increment q;
Output: subset $I_{\boldsymbol{\upsilon}^{(q)}} \subseteq [N]$;

The implementation of Algorithm 4.1 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 4.3. Once $\mathfrak{g}$ is known, each iteration of Algorithm 4.1 has a $\mathcal{O}(N)$ time-complexity. For $q \in \mathbb{N}$, we for instance have

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

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. Assuming that the entries of $\mathbf{S}$ can be accessed on demand, the space-complexity of Algorithm 4.1 is $\mathcal{O}(N)$.

In view of (4.10), the sequence of subsets $I_{\boldsymbol{\upsilon}^{(1)}} \subseteq I_{\boldsymbol{\upsilon}^{(2)}} \subseteq \dots$ generated by Algorithm 4.1 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. Interestingly, taking $\mathbf{f} = \mathfrak{g}$ turns out to be a poor choice in practice. A variant of Algorithm 4.1 with an alternative descent direction is described in Remark 4.2.

Remark 4.2 (Best-improvement direction)**.** Instead of considering the steepest conditional descent directions (4.10), we can combine the information provided by (4.4) and (4.7) to characterise the conditional descent directions inducing the best one-step-ahead improvements. In Algorithm 4.1, we may hence replace the FW

direction (4.10) 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 4.1 is still $\mathcal{O}(N)$; however, in comparison to FW, the resulting procedure is costlier as it additionally requires the computation of the relevant improvement scores. ◁

Remark 4.3 (Enforcing the selection of new columns)**.** In Algorithm 4.1, at step $q \in \mathbb{N}$, the FW direction (4.10) might lead to the selection of a column which already belongs to the sample, that is, we could have $u \in I_{\boldsymbol{\upsilon}^{(q)}}$; we refer to such an event as a *correction step*, since instead of adding a new column to the sample, we are "correcting" the weight associated with an existing column in the sample. To enforce the selection of a new column at each iteration, we may replace the FW direction (4.10) by

$$u \in \arg\min_{i\in[N]}\{\Theta(\boldsymbol{\upsilon}^{(q)};\boldsymbol{\xi}_i)\,|\,i\notin I_{\boldsymbol{\upsilon}^{(q)}} \text{ and } \Theta(\boldsymbol{\upsilon}^{(q)};\boldsymbol{\xi}_i)<0\}; \tag{4.11}$$

if the set characterising (4.11) is empty, the sampling should stop (or an alternative direction should be considered). Such a variant of Algorithm 4.1 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 4.3. ◁

Remark 4.4 (Weight optimisation)**.** For a subset $I \subseteq [N]$ of size $m \leqslant N$, let $\tilde{\boldsymbol{\upsilon}}(I) \in \mathbb{R}^N_{\geqslant 0}$ be a vector minimising D over the set of all selection vectors $\boldsymbol{\upsilon} \in \mathbb{R}^N_{\geqslant 0}$ such that $I_{\boldsymbol{\upsilon}} \subseteq I$ (the nonnegativity of the entries of the PSD matrix $\mathbf{S}$ ensures that such a vector always exists). The non-trivial entries $[\tilde{\boldsymbol{\upsilon}}(I)]_I$ of $\tilde{\boldsymbol{\upsilon}}(I)$ are provided by solutions to the quadratic program (QP) associated with the minimisation of the function $\boldsymbol{x} \mapsto \boldsymbol{x}^*\mathbf{S}_{I,I}\boldsymbol{x} - 2\mathbf{g}_I^*\boldsymbol{x}$ over $\mathbb{R}^m_{\geqslant 0}$. The rescaled vector $\boldsymbol{\upsilon}(I) = \varkappa\tilde{\boldsymbol{\upsilon}}(I)/(\mathbf{f}^*\tilde{\boldsymbol{\upsilon}}(I)) \in \mathcal{A}_{\mathbf{f}}$ then minimises R over the set of all selection vectors $\boldsymbol{\upsilon} \in \mathcal{A}_{\mathbf{f}}$ such that $I_{\boldsymbol{\upsilon}} \subseteq I$. In Algorithm 4.1 and its BI variant, at iteration $q \in \mathbb{N}$, after selection a descent direction u, rather than performing a descent with optimal step size, we may instead set $\boldsymbol{\upsilon}^{(q+1)} = \boldsymbol{\upsilon}(I_{\boldsymbol{\upsilon}^{(q)}} \cup \{u\})$. 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, $\tilde{\boldsymbol{\upsilon}}^{(q)}$ can 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 $I_{\boldsymbol{\upsilon}^{(q)}} \cup \{u\}$; this situation occurs when some entries of the solution to the underlying QP are zero. In the experiments of Section 4.4, instead of the true support $I_{\boldsymbol{\upsilon}^{(q+1)}}$, we keep track of the *virtual support* $\tilde{I}_{\boldsymbol{\upsilon}^{(q+1)}} = \tilde{I}_{\boldsymbol{\upsilon}^{(q)}} \cup \{u\}$, so that $|\tilde{I}_{\boldsymbol{\upsilon}^{(q)}}| = q$ 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 stage of the optimisation process). Enforcing the WO rule on the BI variant of Algorithm 4.1 described in Remark 4.2 leads to a procedure that produces sequences of subsets that are independent of the choice of the restriction vector $\mathbf{f}$; in the framework of Lemma 4.3, we may indeed observe that $\mathcal{I}(\boldsymbol{\upsilon};\boldsymbol{\eta}) = \mathcal{I}(c\boldsymbol{\upsilon};\tilde{c}\boldsymbol{\eta})$ for $c, \tilde{c} > 0$. ◁

4.3 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 approach consists in relying on numerically affordable stochastic approximations of $\mathfrak{g}$. Many approaches could 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 $[N] \setminus \{i\}$. 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

$$\hat{\mathfrak{g}}_{\mathbf{F}} = \operatorname{diag}(\mathbf{S}) + \tfrac{(N-1)}{\ell}(\mathbf{S} \odot \mathbf{F})\mathbb{1}, \tag{4.12}$$

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 S. In the framework of (4.12) 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, let $i, j \in [N]$ with $i \neq j$, and suppose that $i \in \mathcal{S}_j$, that is, suppose that the term $\mathbf{S}_{j,\,i}$ appears in the approximation of $\mathfrak{g}_j$, the j-th entry of $\mathfrak{g}$. We may notice that the same term $\mathbf{S}_{i,\,j} = \mathbf{S}_{j,\,i}$ can 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 $\{\mathbf{F}^*_{i,\,j}\}_{j\in[N]\setminus\{i\}}$ 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 4.5 below), 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 following unbiased estimator of $\mathfrak{g}$:

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

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.

Lemma 4.5 below gives a statistical result that justifies the expected value of the (i, j) entry of the matrix $\mathbf{F}^*$ conditioned on l_i, the number of additional entries included in the approximation of $\mathfrak{g}_i$ when accounting for symmetry.

Lemma 4.5. Let X and Y be two independent random variables following binomial distributions with size parameters m and $n \in \mathbb{N}$, respectively, and with the 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. □

Remark 4.5. Computing a realisation of $\hat{\mathbf{g}}_{\mathbf{F}}$, $\hat{\mathbf{g}}_{\mathbf{F}^*}$ or $\hat{\mathbf{g}}_{\mathrm{sym}}$ involves sampling $(\ell+1)N$ entries of $\mathbf{S}$, with in practice $\ell \ll N$. If ℓ is chosen independently of N, the time-complexity of forming such approximations is linear in N (here, we assume that the complexity of the considered random generator does not depend on N). Assuming that the entries of $\mathbf{S}$ can be accessed on demand, the space-complexity of forming such approximations is also linear in N, and the computation can in addition be easily parallelised. ◁

Sampling driven by an approximate potential. In (4.2) and (4.3), substituting $\mathbf{g}$ with an approximation $\hat{\mathbf{g}} \in \mathbb{R}^N_{\geqslant 0} \setminus \{0\}$ gives rise to approximate error maps $\hat{D}$ and $\hat{R}$. Let $\hat{\mathbb{1}} \in \mathbb{R}^N_{\geqslant 0} \setminus \{0\}$ be a vector minimising $\hat{D}$ over $\mathbb{R}^N_{\geqslant 0}$ (the nonnegativity of the entries of the PSD matrix $\mathbf{S}$ ensures that such a vector always exists). When $\mathbf{g}$ is replaced by $\hat{\mathbf{g}}$, Algorithm 4.1 produces a sequence of selection vectors with increasing supports converging to 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 $\mathbf{g}$ is used throughout the optimisation process (alternative strategies, where the approximation is updated during the optimisation process, could be considered).

Remark 4.6. When a realisation of $\hat{\mathbf{g}}_{\mathbf{F}}$ or $\hat{\mathbf{g}}_{\mathrm{sym}}$ is considered, for $\ell \ll N$, our experiments suggest that the underlying vector $\hat{\mathbb{1}}$ 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 4.4 for illustrations. Following Remark 4.3, for the stochastic variant of Algorithm 4.1, we also observe that considering the modified FW direction (4.11) 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 $\hat{\mathbf{g}}_{\mathbf{F}}$, the reduced variance of the estimator $\hat{\mathbf{g}}_{\mathrm{sym}}$ appears to have a beneficial impact on the column-sampling process. ◁

4.4 Numerical experiments

We now illustrate the behaviour of Algorithm 4.1 and its variants on a series of examples. To assess the accuracy of the Nyström approximation induced by a subset $I \subseteq [N]$ of size $m \leqslant N$, we consider the *approximation factors* (see e.g. Derezinski et al. (2020)).

$$\begin{aligned} \mathscr{E}_{\mathrm{P}}(I) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{HS}(\mathcal{H})}}{\|\mathbf{K} - \mathbf{K}^{\star}_m\|_{\mathrm{HS}(\mathcal{H})}}, \quad \mathscr{E}_{\mathrm{PP}}(I) = \frac{\|\mathbf{K} - \mathbf{P}_I \mathbf{K} \mathbf{P}_I\|_{\mathrm{HS}(\mathcal{H})}}{\|\mathbf{K} - \mathbf{K}^{\star}_m\|_{\mathrm{HS}(\mathcal{H})}}, \\ \text{and} \quad \mathscr{E}_{\mathrm{X}}(I) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(I)\|_{\mathrm{X}}}{\|\mathbf{K} - \mathbf{K}^{\star}_m\|_{\mathrm{X}}}, \ \mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}\}, \end{aligned} \tag{4.13}$$

where $\mathbf{K}^{\star}_m$ is an optimal rank-m approximation of $\mathbf{K}$ (that is, an approximation obtained by spectral truncation; see Section 2.1.2). The values of the approximation factors are necessarily larger than or equal to 1, and the smaller the value, the more accurate the approximation.

Remark 4.7. Denoting by $\lambda_1 \geqslant \ldots \geqslant \lambda_N \geqslant 0$ 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 $\|\mathbf{K} - \mathbf{K}^{\star}_m\|_{\mathrm{sp}} = \lambda_{m+1}$ (see Remark 2.2). ◁

We implement Algorithm 4.1 (referred to as FW, for short) and its BI variant (referred to as BI); see Remark 4.2. 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 Remark 4.4. In the stochastic case, that is, when stochastic approximations of $\mathfrak{g}$ are considered (see Section 4.3), we rely on the estimator $\hat{\mathbf{g}}_{\mathrm{sym}}$ and implement the modified FW direction (4.11); 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 square of the diagonal of $\mathbf{K}$, *leverage-score-based* random sampling, and *determinantal-point-process-based* (DPP-based) random sampling (see Section 2.2.3).

4.4.1 Random PSD matrix

We consider a random PSD matrix $\mathbf{K} \in \mathbb{C}^{N \times N}$, with $N = 1{,}500$; the eigenvalues of $\mathbf{K}$ are independent realisations of a log-normal distribution ($\mu = -2.5$ and $\sigma = 3$), and a set of associated eigenvectors is defined using a random unitary matrix (multiplication-invariant Haar measure; see Mezzadri, 2007). 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 first 100 iterations of Algorithm 4.1 and its BI variant is illustrated in Figure 4.2 (these four error maps are considered since they take the same value at $\boldsymbol{\upsilon} = 0$).

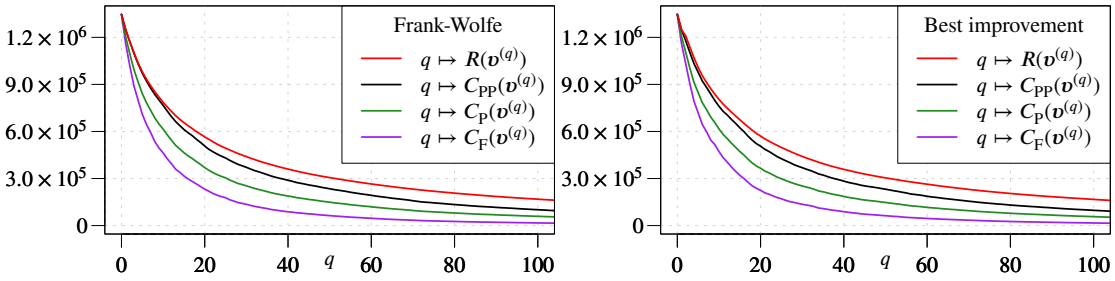

Figure 4.2: For the random PSD matrix example of Section 4.4.1, evolution of the value of the error maps R and C_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}, \mathrm{PP}\}$, during the first 100 iterations of Algorithm 4.1 (left) and its BI variant (right). The exact target potential $\mathfrak{g}$ is used.

In accordance with Lemma 4.4, we see that the error maps C_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}, \mathrm{PP}\}$ are bounded by R throughout. We observe a strong similarity between the evolution of these maps, further supporting the use of R as surrogate error map for Nyström approximation.

We then 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

sampled columns). For the stochastic strategies, 100 repetitions are performed. The results are presented in Figure 4.3. In the considered regime (that is, $m \ll N$), and for all the approximation factors, we observe that the Nyström approximations induced by Algorithm 4.1 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.

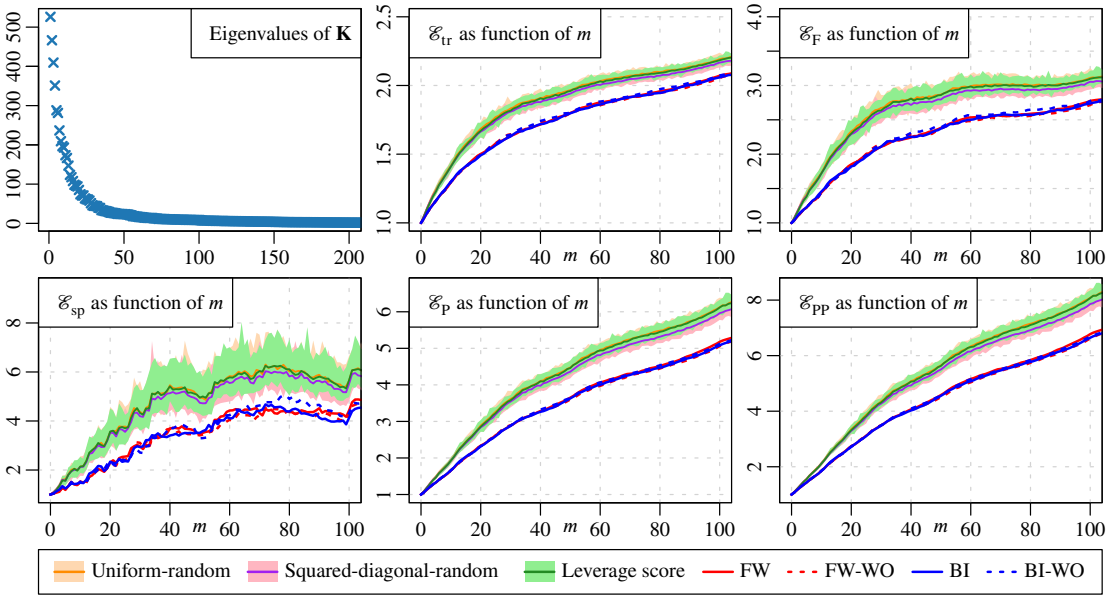

Figure 4.3: For the random PSD matrix example of Section 4.4.1, and for various sampling strategies, evolution of the five approximation factors (4.13) as functions of the number of sampled columns m. The 200 largest eigenvalues of $\mathbf{K}$ are also displayed. For the stochastic methods, the solid line represents the median over 100 repetitions, and the boundaries of the shaded regions indicate the corresponding maximum and minimum values. For the four considered variants of Algorithm 4.1, the exact target potential $\mathfrak{g}$ is used.

4.4.2 Abalone data set

We consider the Abalone data set (UCI Machine Learning Repository; see Dua and Graff, 2019). 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 Gaussian kernel with kernel parameter $\gamma > 0$, given by $K(x, y) = e^{-\gamma \|x-y\|_2^2}$, $x, y \in \mathbb{R}^d$, to generate a PSD matrix $\mathbf{K}$. To illustrate the impact of the decay of the spectrum of $\mathbf{K}$ on the sampling process, we

consider different values of γ, namely $\gamma = 0.1$, 0.25 and 1, chosen so that the the eigenvalues of $\mathbf{K}$ exhibit relatively steep, moderate and shallow decays, respectively; see Figure 4.4.

4.4.2.1 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 4.1 (namely FW, BI, FW-WO and 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 4.4, where we display the evolution of the approximation factors $\mathscr{E}_{\mathrm{F}}$ and $\mathscr{E}_{\mathrm{P}}$ up to $m = 100$ (the evolution of the other approximation factors is provided in Figure 4.7 in the appendix of this chapter; in terms of behaviour, $\mathscr{E}_{\mathrm{tr}}$ and $\mathscr{E}_{\mathrm{sp}}$ appear closely related to $\mathscr{E}_{\mathrm{F}}$, while $\mathscr{E}_{\mathrm{PP}}$ shows similarities with $\mathscr{E}_{\mathrm{P}}$).

Remark 4.8. Following Remark 4.7, in Figure 4.4 (and in the complementary Figure 4.7 in appendix), to illustrate the decay of the spectrum of $\mathbf{K}$ we indicate the thresholds

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

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

In comparison to the considered random-sampling procedures, we observe that Algorithm 4.1 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}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{F}, \mathrm{sp}, \mathrm{P}, \mathrm{PP}\}$, defined in Remark 4.8). After a certain number of iterations, which appears to be related to the decay of the spectrum of $\mathbf{K}$, the relative accuracy of the approximations induced by Algorithm 4.1 and its BI variant deteriorates; this is especially visible for $\gamma = 0.1$. The deterioration is stronger for $\mathscr{E}_{\mathrm{F}}$, $\mathscr{E}_{\mathrm{tr}}$ and $\mathscr{E}_{\mathrm{sp}}$ than for $\mathscr{E}_{\mathrm{P}}$ and $\mathscr{E}_{\mathrm{PP}}$, and the WO update rule appears to be able to mitigate this drop-off in accuracy. Following Lemma 4.4, we recall

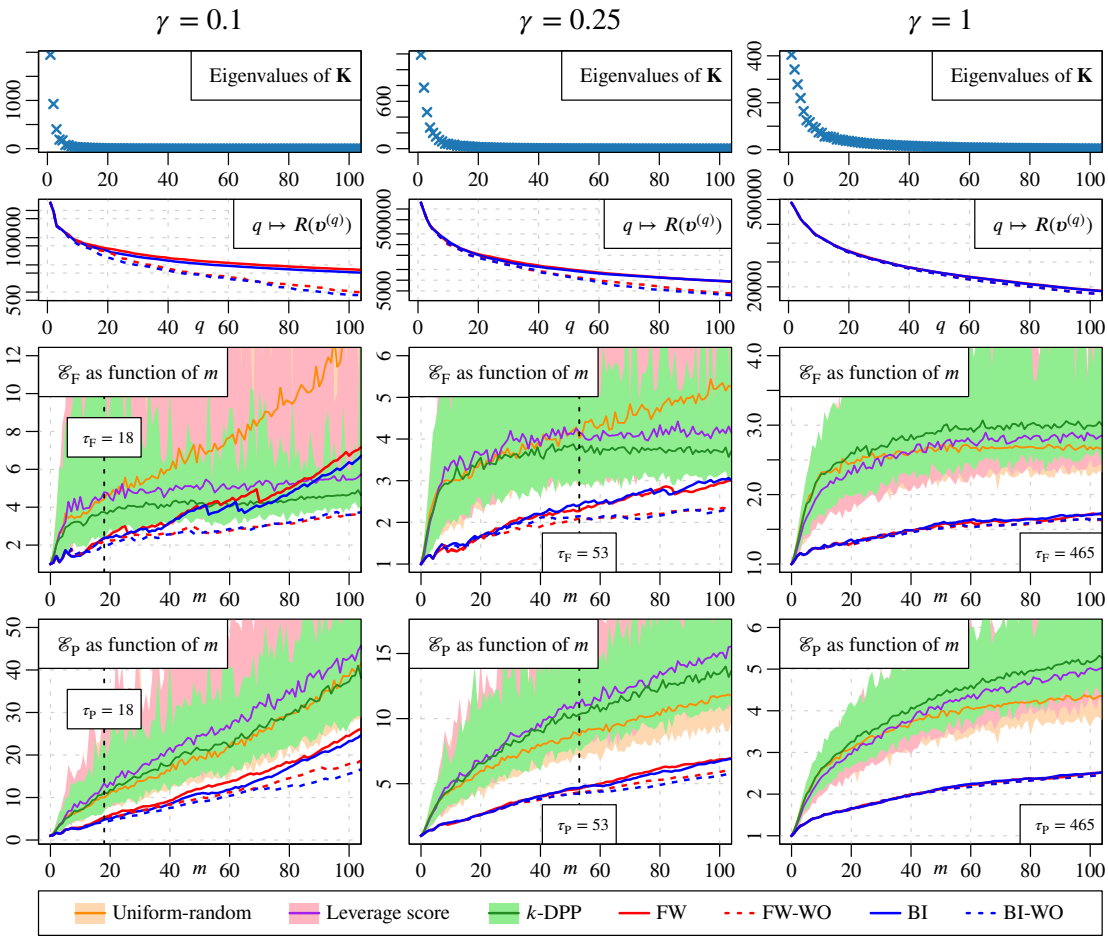

Figure 4.4: For the Abalone data set example of Section 4.4.2, evolution of the approximation factors $\mathscr{E}_{\mathrm{F}}$ and $\mathscr{E}_{\mathrm{P}}$ as functions of the number of sampled columns m (the evolution of the other approximation factors in provided in Figure 4.7 in appendix). Each 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 first 100 iterations of the FW and BI variants of Algorithm 4.1, with both optimal-step-size and WO update rules (the exact target potential $\mathbf{g}$ is used). The evolutions of $\mathscr{E}_{\mathrm{F}}$ and $\mathscr{E}_{\mathrm{P}}$ are represented for the four considered variants of Algorithm 4.1, 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 4.3). The vertical dashed lines indicate the value of the thresholds τ_{X}, $\mathrm{X} \in \{\mathrm{F}, \mathrm{P}\}$, defined in Remark 4.8 (when the threshold is outside the plot window, we only report its value).

that among the considered error maps, C_{P} and C_{PP} are the ones that are the most closely related to R.

4.4.2.2 Approximate target potential

We now consider the stochastic variant S-MFW of Algorithm 4.1, that is, we use realisations of the estimator $\hat{\mathbf{g}}_{\mathrm{sym}}$ (see Section 4.3) in combination with the

modified FW direction (4.11), 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 4.4) and we consider three different values of ℓ, namely $\ell = 100$, 250 and 500. The results are presented in Figure 4.5.

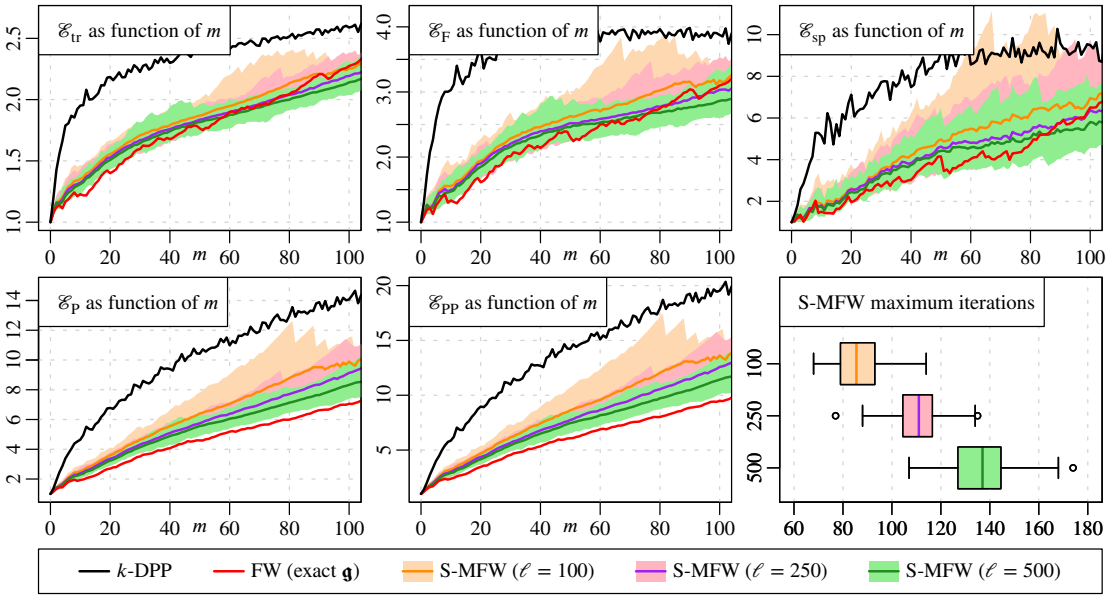

Figure 4.5: For the Abalone data set example of Section 4.4.2 with kernel parameter $\gamma = 0.25$, evolution of the five approximation factors (4.13) as functions of the number of sampled columns m, for samples obtained using the S-MFW variant of Algorithm 4.1 (modified FW direction with realisations of $\hat{\mathfrak{g}}_{\mathrm{sym}}$; see Remark 4.3 and Section 4.3). 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 4.1 (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 4.3).

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 4.3 and 4.6, the maximum number of iterations of the S-MFW procedure tends to increase with ℓ. For this particular example, considering $\ell = 500$ allows for a consistent exploration of the range $m \leqslant 100$ (see Section 4.4.3 for a further illustration of the link between ℓ and the maximum number of S-MFW iterations).

4.4.3 HIGGS data set

We now illustrate the ability of the proposed approach to handle large PSD matrices. We consider the HIGGS dataset (UCI Machine Learning Repository; see Dua and Graff, 2019), 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 Gaussian kernel (same expression as in Section 4.4.2) with $\gamma = 0.1$. In double-precision floating-point format, storing all the entries of $\mathbf{K}$ or $\mathbf{S}$ would require more than 968 terabytes of memory; as an alternative, rather than being stored, the entries of the matrix $\mathbf{S}$ are computed on demand from the data set and the kernel (*on-the-fly evaluation*).

In Figure 4.6, we display the decay of the error map R during the first 50,000 iterations of Algorithm 4.1 (exact target potential). Lemma 4.4 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 4.2 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 $\mathbf{K}$.

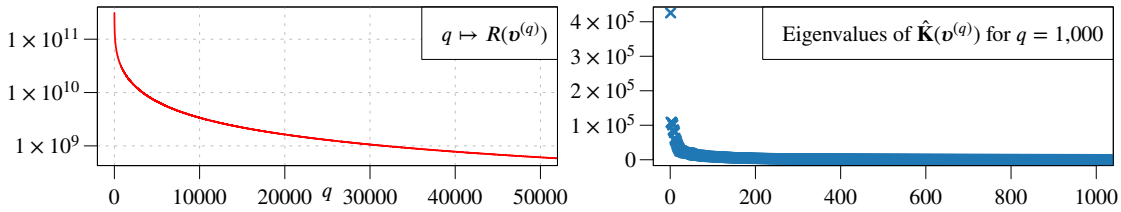

Figure 4.6: For the HIGGS data set example of Section 4.4.3, decay of the the error map R during the first 50,000 iterations of Algorithm 4.1 (logarithmic scale). The non-zero eigenvalues of the Nyström approximation of $\mathbf{K}$ obtained at $q = 1{,}000$ are also presented.

We next implement the S-MFW variant of Algorithm 4.1 for 10 realisations of the estimator $\hat{\mathbf{g}}_{\mathrm{sym}}$ 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 4.3). 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 4.1.

As observed in Sections 4.4.1 and 4.4.2, the samples of columns obtained using Algorithm 4.1 and its S-MFW stochastic variant are noticeably more accurate than

Table 4.1: 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 4.1 with $\ell = 10{,}000$ (stochastic approximations of $\mathfrak{g}$), as well as for the deterministic column samples produced by Algorithm 4.1 (exact target potential $\mathfrak{g}$).

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	—

the ones obtained through random uniform 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, $N/\ell = 1{,}100$).

Appendix: Additional figure for Abalone data set

Figure 4.7 complements Figure 4.4 by providing the evolution, as functions of the number of sampled 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 4.4.2.1 involving the exact target potential $\mathfrak{g}$.

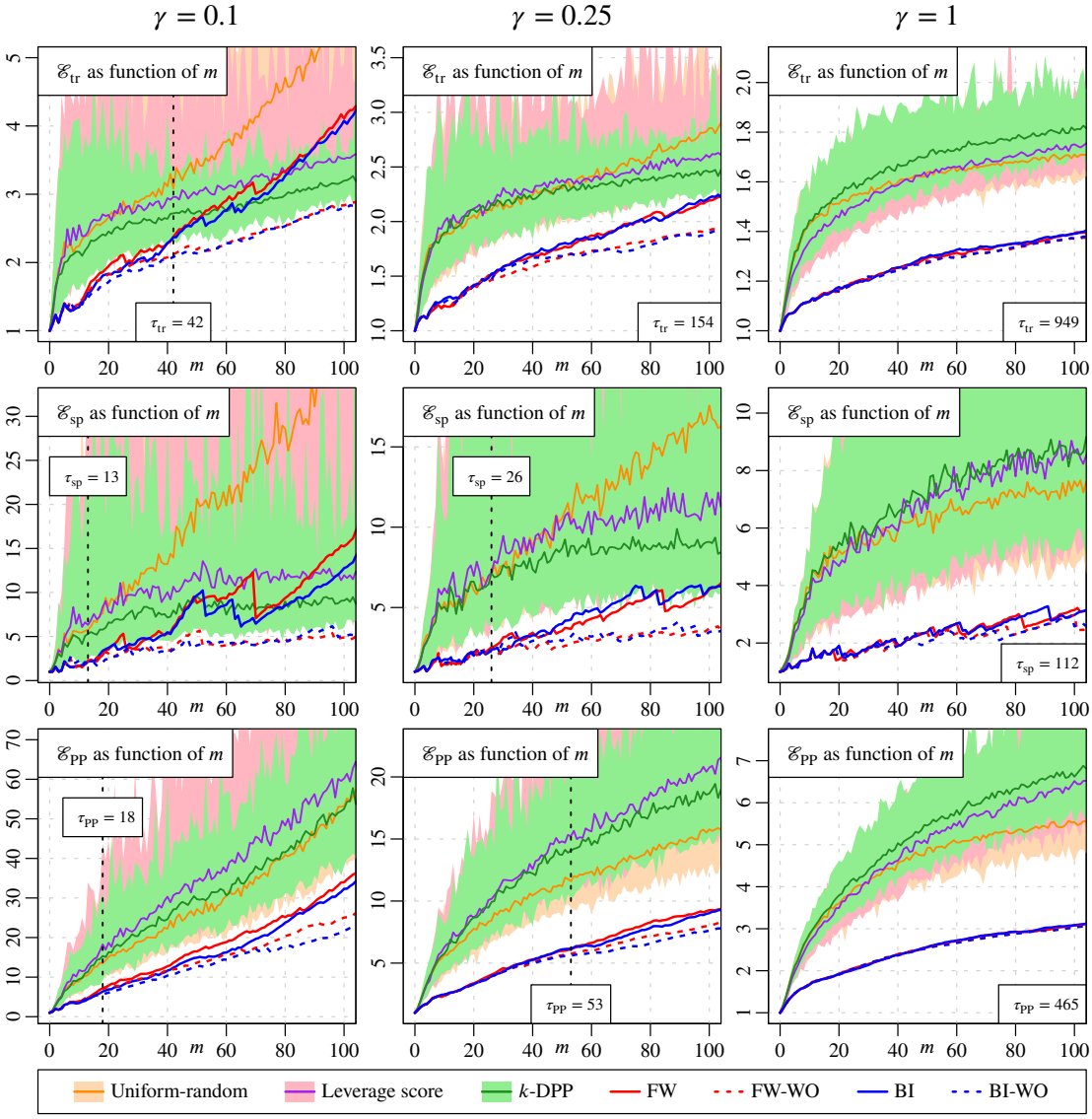

Figure 4.7: For the Abalone data set example of Section 4.4.2, and in complement to Figure 4.4, evolution of the approximation factors $\mathscr{E}_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{sp}, \mathrm{PP}\}$, as functions of the number of sampled columns m for the various sampling strategies considered in Section 4.4.2.1; the values of the corresponding thresholds τ_{X}, $\mathrm{X} \in \{\mathrm{tr}, \mathrm{sp}, \mathrm{PP}\}$, are also indicated (see Remark 4.8).

Chapter 5

Particle flow-based sampling strategies

In this chapter, we study a relaxed version of the column-sampling problem (CSP) for the Nyström approximation of kernel matrices, where approximations are defined from multisets of landmark points in the ambient space.

The chapter is organised as follows. In Section 5.1, we describe the considered landmark-point relaxation of the CSP, and we define the trace, Frobenius and spectral error maps in this setting; in Section 5.2, we introduce an energy-based differentiable map $\mathfrak{R}$ that is used as a surrogate for these classical error maps. Section 5.3 consists of Theorem 5.1, the main theoretical result of this chapter, which guarantees the convergence of gradient descent iterates of $\mathfrak{R}$ under reasonable assumptions. A discussion of stochastic approximations of the gradient of $\mathfrak{R}$ is contained in Section 5.4, and we present a number of numerical experiments in Section 5.5. The appendix of this chapter contains a proof of Theorem 5.1.

In this chapter, we for simplicity only consider real symmetric positive semidefinite (SPSD) matrices.

5.1 Approximation of PSD kernel matrices

Let $\mathscr{X}$ be a general space, and let $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be a real-valued PSD kernel function (popular PSD choices for K include the Gaussian kernel and the Matérn kernels; see e.g. Rasmussen and Williams, 2006). A kernel function and a multiset $\mathcal{D} = \{x_1, \ldots, x_N\} \subseteq \mathscr{X}$ define an SPSD kernel matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ with (i, j) entry $K(x_i, x_j)$, $i, j \in [N]$.

5.1.1 Nyström approximation through landmark points

Consider the CSP as discussed in Section 2.2.2, and let $\mathbf{K}$ be a kernel matrix defined from a PSD kernel $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ and a multiset $\mathcal{D} = \{x_1, \dots, x_N\} \subseteq \mathscr{X}$. A sample of columns from $\mathbf{K}$ is thus naturally associated with a subset of $\mathcal{D}$; more precisely, a sample of $m \in \mathbb{N}$ columns of $\mathbf{K}$, indexed by $I = \{i_1, \dots, i_m\} \subseteq [N]$, defines a multiset $\{x_{i_1}, \dots, x_{i_m}\} \subseteq \mathcal{D}$, so that the induced Nyström approximation may be regarded as an approximation induced by a subset of points in $\mathcal{D}$. Consequently, in the kernel-matrix setting, instead of relying only on subsets of columns, we can more generally consider Nyström approximations defined from a multiset $\mathcal{S} \subseteq \mathscr{X}$. Using matrix notation, the Nyström approximation of $\mathbf{K}$ defined by a subset $\mathcal{S} = \{s_1, \dots, s_m\}$ is the $N \times N$ SPSD matrix $\hat{\mathbf{K}}(\mathcal{S})$, with (i, j) entry

$$\left[\hat{\mathbf{K}}(\mathcal{S})\right]_{i,\,j} = \mathbf{k}_{\mathcal{S}}^T(x_i)\mathbf{K}_{\mathcal{S}}^{\dagger}\mathbf{k}_{\mathcal{S}}(x_j), \tag{5.1}$$

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

$$\mathbf{k}_{\mathcal{S}}(x) = \Big(K(x, s_1), \dots, K(x, s_m)\Big)^T \in \mathbb{R}^m.$$

Throughout this chapter, we use the terminology "Nyström sample" to refer to a set or multiset $\mathcal{S} \subseteq \mathscr{X}$, as opposed to a column sample $I \subseteq N$ of the matrix $\mathbf{K}$ discussed in previous chapters. We call the elements of $\mathcal{S}$ *landmark points* (the terminology *inducing points* can also be found in the literature, see e.g. Meanti et al., 2022); the notation $\hat{\mathbf{K}}(\mathcal{S})$ emphasises that the considered Nyström approximation of $\mathbf{K}$ is induced by $\mathcal{S}$.

Remark 5.1. Denoting by $\mathcal{H}$ the RKHS of real-valued functions on $\mathscr{X}$ with reproducing kernel K, we note that the matrix $\hat{\mathbf{K}}(\mathcal{S})$ is the kernel matrix defined by K_S and $\mathcal{D}$, with K_S the reproducing kernel of the closed linear subspace

$$\mathcal{H}_S = \operatorname{span}\{k_{s_1}, \dots, k_{s_m}\} \subseteq \mathcal{H},$$

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

Remark 5.2. When equipped with weights $\{v_i\}_{i=1}^m \in \mathbb{R}^m$, the landmark points $\{s_i\}_{i=1}^m \subseteq \mathscr{X}$ may be regarded as *particles* in the context of particle flow-based

optimisation techniques, in which a target measure μ is approximated by a discrete measure ν which is a weighted sum of Dirac measures at points in $\mathscr{X}$, that is,

$$\nu = \sum_{i=1}^{m} \upsilon_i \delta_{s_i}.$$

The particle flow formulation has connections to optimal transport theory (see e.g. Santambrogio (2015)), and gradient descent techniques performed on the weight and position parameters of the particles have been developed, with for instance applications in the training of neural networks (see e.g. Chizat and Bach (2018)). Particle-flow-based techniques have also been applied to the problem of batch Bayesian optimisation (see e.g. Crovini et al. (2022)). ◁

As in the column-sampling case, the landmark point framework naturally raises questions related to the efficient characterisation and design of accurate Nyström samples (i.e. samples leading to accurate approximations of $\mathbf{K}$).

In this chapter, for a fixed $m \in \mathbb{N}$, we interpret Nyström samples of size m as elements of $\mathscr{X}^m$, and we investigate the possibility of directly optimising Nyström samples over $\mathscr{X}^m$. We consider the case $\mathscr{X} = \mathbb{R}^d$, but $\mathscr{X}$ may more generally be a differentiable manifold.

5.1.2 Approximation accuracy

For a Nyström sample $\mathcal{S} \subseteq \mathscr{X}$ of size $m \leqslant N$, the accuracy of the induced Nyström approximation $\hat{\mathbf{K}}(\mathcal{S})$ can be assessed through the following error maps, which are based on the trace, Frobenius and spectral norms of the approximation error:

(C.1) $\mathfrak{C}_{\mathrm{tr}}(\mathcal{S}) = \left\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\right\|_{\mathrm{tr}}$;

(C.2) $\mathfrak{C}_{\mathrm{F}}(\mathcal{S}) = \left\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\right\|_{\mathrm{F}}$;

(C.3) $\mathfrak{C}_{\mathrm{sp}}(\mathcal{S}) = \left\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\right\|_{\mathrm{sp}}$.

As discussed in Section 2.1.1, the computation of the trace, Frobenius and spectral norm errors becomes prohibitively expensive when N is large (see Table 2.1). Additionally, the evaluation of the partial derivatives of these maps (regarded as functions from $\mathscr{X}^m = \mathbb{R}^{md}$ to $\mathbb{R}$) with respect to a single coordinate of a landmark point has a complexity similar to the complexity of evaluating the maps themselves (and there are in this case md such partial derivatives). Consequently, a direct optimisation of these maps over $\mathscr{X}^m$ is intractable in most practical applications.

5.2 An energy-based error map

As a surrogate for the error maps (C.1)-(C.3), for $\mathcal{S} = \{s_1, \ldots, s_m\}$, we introduce the map

$$\mathfrak{R}(\mathcal{S}) = \begin{cases} \|\mathbf{K}\|_{\mathrm{F}}^2 - \dfrac{1}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2} \left(\displaystyle\sum_{i=1}^{N} \sum_{j=1}^{m} K^2(x_i, s_j) \right)^2, & \text{if } \|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}} > 0, \\ \|\mathbf{K}\|_{\mathrm{F}}^2, & \text{otherwise,} \end{cases} \tag{5.2}$$

where $K^2(x_i, s_j)$ stands for $\Big(K(x_i, s_j)\Big)^2$. We may note that $0 \leqslant \mathfrak{R}(\mathcal{S}) \leqslant \|\mathbf{K}\|_{\mathrm{F}}^2$.

In (5.2), the evaluation of the term $\|\mathbf{K}\|_{\mathrm{F}}^2$ has complexity $\mathcal{O}(N^2)$; nevertheless, this term does not depend on the Nyström sample $\mathcal{S}$, and may thus be regarded as a constant. The complexity of the evaluation of the term $\mathfrak{R}(\mathcal{S}) - \|\mathbf{K}\|_{\mathrm{F}}^2$, that is, of evaluating $\mathfrak{R}(\mathcal{S})$ up to the constant $\|\mathbf{K}\|_{\mathrm{F}}^2$, is $\mathcal{O}(m^2 + mN)$; for $\mathscr{X} = \mathbb{R}^d$, the same holds for the complexity of the evaluation of the partial derivative of $\mathfrak{R}(\mathcal{S})$ with respect to a coordinate of a landmark point (see equation (5.4) below). Importantly, and in contrast to the error maps discussed in Section 5.1.2, the evaluation of $\mathfrak{R}(\mathcal{S})$ or of its partial derivatives does not involve the inversion or pseudoinversion of the $m \times m$ matrix $\mathbf{K}_{\mathcal{S}}$.

Remark 5.3. From a theoretical standpoint, the map $\mathfrak{R}$ arises as a scaling-invariant (see Section 4.1.3) version of the square of the MMD described in Section 3.1.3.2 when interpreted in the landmark-point setting. In this case, the measure μ is taken to be the uniform measure on $\mathcal{D}$, that is

$$\mu = \sum_{i=1}^{N} \delta_{x_i}.$$

The Nyström sample $\mathcal{S} = \{s_1, \ldots, s_m\} \subseteq \mathscr{X} = \mathbb{R}^d$ is regarded as a set of landmark points, and we take ν to be the uniform measure on $\mathcal{S}$, that is,

$$\nu = \sum_{j=1}^{m} \delta_{s_j}.$$

The map $\mathfrak{R}$ may also be defined for non-uniform measures ν, and in this case depends not only on $\mathcal{S}$, but also on a set of relative weights associated with each landmark point in $\mathcal{S}$. In this chapter, we for simplicity only focus on the uniform case. ◁

The following inequalities hold:

$$\left\|\mathbf{K}-\hat{\mathbf{K}}(\mathcal{S})\right\|_{\mathrm{sp}}^2 \leqslant \left\|\mathbf{K}-\hat{\mathbf{K}}(\mathcal{S})\right\|_{\mathrm{F}}^2 \leqslant \mathfrak{R}(\mathcal{S}) \leqslant \|\mathbf{K}\|_{\mathrm{F}}^2,$$

and

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

which, in complement to the theoretical properties enjoyed by the error map $\mathfrak{R}$, further support the use of $\mathfrak{R}$ as a numerically-affordable surrogate for (C.1)-(C.3) (see also the numerical experiments in Section 5.5).

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

$$\partial^{[\mathrm{l}]}_{[s]_l} K^2(s,x) \quad \text{and} \quad \partial^{[\mathrm{d}]}_{[s]_l} K^2(s,s) \tag{5.3}$$

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

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

$$\begin{aligned}\partial_{[s_k]_l}\mathfrak{R}(\mathcal{S}) = \frac{1}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^4}\left(\sum_{i=1}^N\sum_{j=1}^m K^2(s_j,x_i)\right)^2\left(\partial^{[\mathrm{d}]}_{[s_k]_l}K^2(s_k,s_k) + 2\sum_{\substack{j=1,\\ j\neq k}}^m \partial^{[\mathrm{l}]}_{[s_k]_l}K^2(s_k,s_j)\right)\\ - \frac{2}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2}\left(\sum_{i=1}^N\sum_{j=1}^m K^2(s_j,x_i)\right)\left(\sum_{i=1}^N \partial^{[\mathrm{l}]}_{[s_k]_l}K^2(s_k,x_i)\right).\end{aligned} \tag{5.4}$$

By mutualising the evaluation of the terms in (5.4) that do not depend on k and l, the evaluation of the md partial derivatives of $\mathfrak{R}$ at $\mathcal{S}$ has a complexity of $\mathcal{O}\big((d+1)(m^2+mN)\big)$; by contrast (and although the pseudoinversion of $\mathbf{K}_{\mathcal{S}}$ can be mutualised), evaluating the md partial derivatives of the trace error map has a complexity of $\mathcal{O}\big(d(m^4+m^3N)\big)$.

In this chapter, we investigate the possibility to use the partial derivatives (5.4), or stochastic approximations of these derivatives, to directly optimise the error map $\mathfrak{R}$ over $\mathscr{X}^m$ via gradient or stochastic gradient descent; the stochastic approximation schemes we consider aim at reducing the burden of the numerical cost induced by the evaluation of the partial derivatives of $\mathfrak{R}$ when N is large.

5.3 A convergence result

We use the same notation as in Section 5.2 (in particular, we still assume that $\mathscr{X} = \mathbb{R}^d$), and by analogy with (5.3), for $s, x \in \mathscr{X}$, and for $l \in [d]$, we denote by $\partial^{[\mathrm{r}]}_{[s]_l} K^2(x, s)$ the partial derivative of the map $s \mapsto K^2(x, s)$ with respect to the l-th coordinate of s. Also, for a fixed $m \in \mathbb{N}$, we denote by $\nabla\mathfrak{R}(\mathcal{S}) \in \mathscr{X}^m = \mathbb{R}^{md}$ the gradient of $\mathfrak{R} : \mathscr{X}^m \to \mathbb{R}$ at $\mathcal{S}$; in matrix notation, we have

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

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

Theorem 5.1. We make the following assumptions on the squared-kernel K^2, which we assume hold for all $x, y \in \mathscr{X} = \mathbb{R}^d$, and all $l, l' \in [d]$, uniformly:

(A.1) there exists $\alpha > 0$ such that for all $x \in \mathbb{R}^d$,

$$K^2(x, x) \geqslant \alpha;$$

(A.2) there exists $M_1 > 0$ such that for all $x, y \in \mathbb{R}^d$ and all $l \in [d]$,

$$\big|\partial^{[\mathrm{d}]}_{[x]_l} K^2(x, x)\big| \leqslant M_1 \quad \text{and} \quad \big|\partial^{[\mathrm{l}]}_{[x]_l} K^2(x, y)\big| \leqslant M_1;$$

(A.3) there exists $M_2 > 0$ such that for all $x, y \in \mathbb{R}^d$ and all $l, l' \in [d]$,

$$\big|\partial^{[\mathrm{d}]}_{[x]_l}\partial^{[\mathrm{d}]}_{[x]_{l'}} K^2(x, x)\big| \leqslant M_2, \quad \big|\partial^{[\mathrm{l}]}_{[x]_l}\partial^{[\mathrm{l}]}_{[x]_{l'}} K^2(x, y)\big| \leqslant M_2,$$
$$\text{and} \quad \big|\partial^{[\mathrm{l}]}_{[x]_l}\partial^{[\mathrm{r}]}_{[y]_{l'}} K^2(x, y)\big| \leqslant M_2.$$

Under the above assumptions, there exists an $L > 0$ such that for any two Nyström samples $\mathcal{S}, \mathcal{S}' \in \mathbb{R}^{md}$,

$$\big\|\nabla\mathfrak{R}(\mathcal{S}) - \nabla\mathfrak{R}(\mathcal{S}')\big\|_2 \leqslant L\big\|\mathcal{S} - \mathcal{S}'\big\|_2,$$

with $\|\cdot\|_2$ the Euclidean norm of $\mathbb{R}^{md}$; in other words, the gradient of $\mathfrak{R} : \mathbb{R}^{md} \to \mathbb{R}$ is Lipschitz-continuous with Lipschitz constant L.

Since $\mathfrak{R}$ is bounded from below, for $0 < \rho \leqslant 1/L$ and independently of the considered initial Nyström sample $\mathcal{S}^{(0)}$, Theorem 5.1 entails that a gradient descent from $\mathcal{S}^{(0)}$, with fixed step size ρ, for the minimisation of $\mathfrak{R}$ over $\mathscr{X}^m$ produces a sequence of iterates that converges to a critical point of $\mathfrak{R}$. Barring some specific and largely pathological cases, the resulting critical point is likely to be a local minimum of $\mathfrak{R}$ (see for instance Lee et al., 2016). A proof of Theorem 5.1 is provided in the appendix of this chapter.

The conditions considered in Theorem 5.1 ensure the existence of a general Lipschitz constant L for the gradient of $\mathfrak{R}$; they, for instance, hold for all sufficiently regular Matérn kernels (thus including the Gaussian kernel). We stress that these conditions are only sufficient conditions for the convergence of a gradient descent for the minimisation of $\mathfrak{R}$; by introducing additional problem-dependent conditions, some convergence results could be obtained for more general kernels K^2 and adequate initial Nyström samples $\mathcal{S}^{(0)}$. For instance, the condition (A.1) simply ensures that $\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2 \geqslant m\alpha > 0$ for all $\mathcal{S} \in \mathscr{X}^m$; this condition could be relaxed to account for kernels with vanishing diagonal, but one might then need to introduce ad hoc conditions to ensure that $\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2$ remains large enough during the minimisation process.

5.4 Stochastic approximation of the gradient

The complexity of evaluating a single partial derivative of $\mathfrak{R} : \mathscr{X}^m \to \mathbb{R}$ is $\mathcal{O}(m^2 + mN)$, which might become prohibitive for large values of N. To overcome this limitation, stochastic approximations of the gradient of $\mathfrak{R}$ can be considered (see e.g. Bottou et al., 2018).

The evaluation of (5.4) involves, for instance, terms of the form $\sum_{i=1}^N K^2(s, x_i)$, with $s \in \mathscr{X}$ and $\mathcal{D} = \{x_1, \ldots, x_N\}$. Introducing a random variable X with a uniform distribution on $\mathcal{D}$, we can observe that

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

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

$$\mathbb{E}\Big[K^2(s,X)\Big] = \frac{1}{b}\sum_{j=1}^{b}\mathbb{E}\Big[K^2(s,X_j)\Big]$$

and

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

so that we can easily define unbiased estimators of the various terms appearing in (5.4). We refer to the sample size b as the *batch size*.

Let $k\in[m]$ and $l\in[d]$; the partial derivative (5.4) can be rewritten as

$$\partial_{[s_k]_l}\mathfrak{R}(\mathcal{S}) = \frac{T_1^2}{\|\mathbf{K}_\mathcal{S}\|_\mathrm{F}^4}\Upsilon(\mathcal{S}) - \frac{2T_1T_2^{k,l}}{\|\mathbf{K}_\mathcal{S}\|_\mathrm{F}^2},$$

with

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

and

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

The terms T_1 and $T_2^{k,l}$ are the only terms in (5.4) that depend on $\mathcal{D}$. From a random sample $\mathbf{X}=\{X_1,\dots,X_b\}$, we define the unbiased estimators $\hat{T}_1(\mathbf{X})$ of T_1, and $\hat{T}_2^{k,l}(\mathbf{X})$ of $T_2^{k,l}$, as

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

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

One-Sample Approximation. Using a single random sample $\mathbf{X}=\{X_1,\dots,X_b\}$ of size b, we can define the following stochastic approximation of the partial derivative (5.4):

$$\hat{\partial}_{[s_k]_l}\mathfrak{R}(\mathcal{S};\mathbf{X}) = \frac{\hat{T}_1(\mathbf{X})^2}{\|\mathbf{K}_\mathcal{S}\|_\mathrm{F}^4}\Upsilon(\mathcal{S}) - \frac{2\hat{T}_1(\mathbf{X})\hat{T}_2^{k,l}(\mathbf{X})}{\|\mathbf{K}_\mathcal{S}\|_\mathrm{F}^2}. \tag{5.5}$$

An evaluation of $\hat{\partial}_{[s_k]_l}\mathfrak{R}(\mathcal{S};\mathbf{X})$ has complexity $\mathcal{O}(m^2+mb)$, as opposed to $\mathcal{O}(m^2+mN)$ for the corresponding exact partial derivative. However, due to

the dependence between $\hat{T}_1(\mathbf{X})$ and $\hat{T}_2^{k,l}(\mathbf{X})$, and to the fact that $\hat{\partial}_{[s_k]_l}\mathfrak{R}(\mathcal{S};\mathbf{X})$ involves the square of $\hat{T}_1(\mathbf{X})$, the stochastic partial derivative $\hat{\partial}_{[s_k]_l}\mathfrak{R}(\mathcal{S};\mathbf{X})$ will generally be a biased estimator of $\partial_{[s_k]_l}\mathfrak{R}(\mathcal{S})$.

Two-Sample Approximation. To obtain an unbiased estimator of the partial derivative (5.4), instead of considering a single random sample, we may define a stochastic approximation based on two independent random samples $\mathbf{X} = \{X_1, \ldots, X_{b_\mathbf{X}}\}$ and $\mathbf{Y} = \{Y_1, \ldots, Y_{b_\mathbf{Y}}\}$, consisting of $b_\mathbf{X}$ and $b_\mathbf{Y} \in \mathbb{N}$ copies of X (i.e. consisting of uniform random variables on $\mathcal{D}$), with $b = b_\mathbf{X} + b_\mathbf{Y}$. The two-sample estimator of (5.4) is then given by

$$\hat{\partial}_{[s_k]_l}\mathfrak{R}(\mathcal{S};\mathbf{X},\mathbf{Y}) = \frac{\hat{T}_1(\mathbf{X})\hat{T}_1(\mathbf{Y})}{\|\mathbf{K}_\mathcal{S}\|_\mathrm{F}^4}\Upsilon(\mathcal{S}) - \frac{2\hat{T}_1(\mathbf{X})\hat{T}_2^{k,l}(\mathbf{Y})}{\|\mathbf{K}_\mathcal{S}\|_\mathrm{F}^2}, \tag{5.6}$$

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

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

Although being unbiased, for a common batch size b, the variance of the two-sample estimator (5.6) will generally be larger than the variance of the one-sample estimator (5.5). In our numerical experiments, the larger variance of the unbiased estimator (5.6) seems to actually slow down the descent when compared to the descent obtained with the one-sample estimator (5.5).

Remark 5.4. While considering two independent samples $\mathbf{X}$ and $\mathbf{Y}$, the two terms $\hat{T}_1(\mathbf{X})\hat{T}_1(\mathbf{Y})$ and $\hat{T}_1(\mathbf{X})\hat{T}_2^{k,l}(\mathbf{Y})$ in (5.6) are dependent. This dependence may complicate the analysis of the properties of the resulting SGD; nevertheless, this issue might be overcome by considering four independent samples instead of two. ◁

5.5 Numerical experiments

Throughout this section, the considered matrices $\mathbf{K}$ are defined from multisets $\mathcal{D} = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ and from Gaussian kernels K with kernel parameter $\gamma > 0$, that is, kernels of the form $K(x,y) = e^{-\gamma\|x-y\|^2}$, $x, y \in \mathbb{R}^d$. Except for the synthetic example of Section 5.5.1, all the multisets $\mathcal{D}$ we consider consist of the entries of data sets available on the UCI Machine Learning Repository (see Dua and Graff, 2019).

Our experiments are based on the following protocol: for a given $m \in \mathbb{N}$, we consider an initial Nyström sample $\mathcal{S}^{(0)}$ consisting of m points drawn uniformly at random, without replacement, from $\mathcal{D}$. The initial sample $\mathcal{S}^{(0)}$ is regarded as an element of $\mathscr{X}^m$, and is used to initialise an SGD (except in Section 5.5.1, where GD is used), with fixed step size $\rho > 0$, for the minimisation of $\mathfrak{R}$ over $\mathscr{X}^m$, yielding, after $T \in \mathbb{N}$ iterations, a locally optimised Nyström sample $\mathcal{S}^{(T)}$. The SGDs are performed with the one-sample estimator (5.5) and are based on independent and identically distributed uniform random variables on $\mathcal{D}$ (i.e. i.i.d. sampling), with batch size $b \in \mathbb{N}$; see Section 5.4.

We assess the accuracy of the Nyström approximations of $\mathbf{K}$ induced by $\mathcal{S}^{(0)}$ and $\mathcal{S}^{(T)}$ in terms of the error map $\mathfrak{R}$ and of the classical error maps (C.1)-(C.3) (for large matrices, we only consider the trace norm). We in parallel investigate the impact of the Nyström sample size (Sections 5.5.1 and 5.5.3) and of the kernel parameter (Section 5.5.2), and demonstrate the ability of the proposed approach to tackle problems of relatively large size (Section 5.5.4).

For a Nyström sample $\mathcal{S} \in \mathscr{X}^m$ of size $m \in \mathbb{N}$, the matrix $\hat{\mathbf{K}}(\mathcal{S})$ is of rank at most m. Following the works (Gittens and Mahoney, 2016; Derezinski et al., 2020), to assess the accuracy of the approximation of $\mathbf{K}$ induced by $\mathcal{S}$, we consider the *approximation factors*

$$\begin{aligned} \mathfrak{E}_{\mathrm{tr}}(\mathcal{S}) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_{\mathrm{tr}}}{\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{tr}}}, \quad \mathfrak{E}_{\mathrm{F}}(\mathcal{S}) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_{\mathrm{F}}}{\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{F}}}, \\ \text{and} \quad \mathfrak{E}_{\mathrm{sp}}(\mathcal{S}) = \frac{\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_{\mathrm{sp}}}{\|\mathbf{K} - \mathbf{K}^\star_m\|_{\mathrm{sp}}}, \end{aligned} \tag{5.7}$$

where $\mathbf{K}^\star_m$ denotes an optimal rank-m approximation of $\mathbf{K}$ (i.e. the approximation of $\mathbf{K}$ obtained by truncation of a spectral expansion of $\mathbf{K}$ and based on m of the largest eigenvalues of $\mathbf{K}$; see Section 2.1.2). The closer $\mathfrak{E}_{\mathrm{tr}}(\mathcal{S})$, $\mathfrak{E}_{\mathrm{F}}(\mathcal{S})$ and $\mathfrak{E}_{\mathrm{sp}}(\mathcal{S})$ are to 1, the more accurate the approximation.

5.5.1 Bi-Gaussian example

We consider a kernel matrix $\mathbf{K}$ defined by a set $\mathcal{D}$ consisting of $N = 2{,}000$ points in $[-1, 1]^2 \subset \mathbb{R}^2$ (i.e. $d = 2$); for the Gaussian kernel parameter, we use $\gamma = 1$. A graphical representation of the set $\mathcal{D}$ is given in Figure 5.1; it consists of N

independent realisations of a bivariate random variable whose density is proportional to the restriction of a bi-Gaussian density to the set $[-1, 1]^2$ (the two modes of the underlying distribution are located at $(-0.8, 0.8)$ and $(0.8, -0.8)$, and the covariance matrix of each Gaussian density is $\mathbf{I}_2/2$, with $\mathbf{I}_2$ the 2×2 identity matrix).

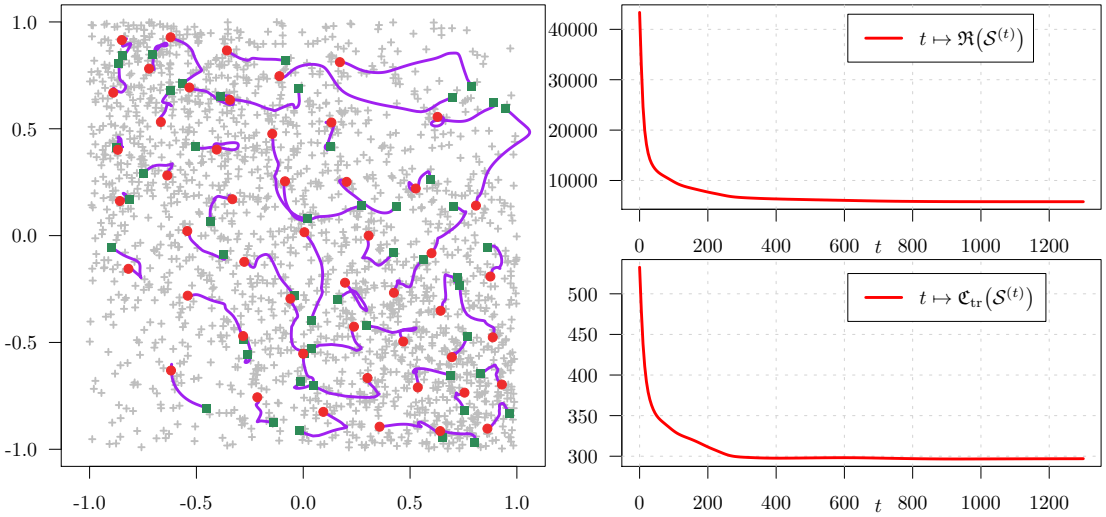

Figure 5.1: For the bi-Gaussian example of Section 5.5.1, graphical representation of the path $t \mapsto \mathcal{S}^{(t)}$ followed by the landmark points of a Nystrom sample during the local minimisation of $\mathfrak{R}$ through GD, with $m = 50$, $\rho = 10^{-6}$ and $T = 1{,}300$; the green squares are the landmark points of the initial sample $\mathcal{S}^{(0)}$, the red dots are the landmark points of the locally optimised sample $\mathcal{S}^{(T)}$, and the purple lines correspond to the paths followed by each landmark point. The grey crosses are the points in $\mathcal{D}$ (left). The evolution, during the GD, of $\mathfrak{R}$ and the trace error map is also presented (right).

The initial samples $\mathcal{S}^{(0)}$ are optimised via GD with step size $\rho = 10^{-6}$ and for a fixed number of iterations T. A graphical representation of the paths followed by the landmark points during the optimisation process is given in Figure 5.1 (for $m = 50$ and $T = 1{,}300$); we observe that the landmark points exhibit a relatively complex dynamic, some of them showing significant displacements from their initial positions. The optimised landmark points concentrate around the regions where the density of points in $\mathcal{D}$ is the largest, and inherit a space-filling-type property in accordance with the stationarity of the kernel K. We also observe that the minimisation of $\mathfrak{R}$ induces a significant decay of the trace error (C.1).

To assess the improvement, in terms of Nyström approximation, yielded by the optimisation of $\mathfrak{R}$, for a given number of landmark points $m \in \mathbb{N}$, we randomly draw an initial Nyström sample $\mathcal{S}^{(0)}$ from $\mathcal{D}$ (uniform sampling without replacement) and compute the corresponding locally optimised sample $\mathcal{S}^{(T)}$ (GD with $\rho = 10^{-6}$ and $T = 1{,}000$). We then compare $\mathfrak{R}\big(\mathcal{S}^{(0)}\big)$ with $\mathfrak{R}\big(\mathcal{S}^{(T)}\big)$, and compute the

corresponding approximation factors with respect to the trace, Frobenius and spectral norms, see (5.7). We consider three different values of m, namely $m = 20$, 50 and 80, and each time perform $n = 1{,}000$ repetitions of this experiment.

Our results are presented in Figure 5.2; we observe that, independently of m, the local optimisation produces a significant improvement in the Nyström approximation accuracy for all the considered error maps; the improvements are particularly noticeable for the trace and Frobenius norms, and slightly less for the spectral norm (which, of the three, appears the coarsest measure of the approximation accuracy). Remarkably, the accuracies of the locally optimised Nyström samples are relatively close to each other, in particular in terms of trace and Frobenius norms, suggesting that a large proportion of the local minima of the error map $\mathfrak{R}$ induce approximations of comparable quality.

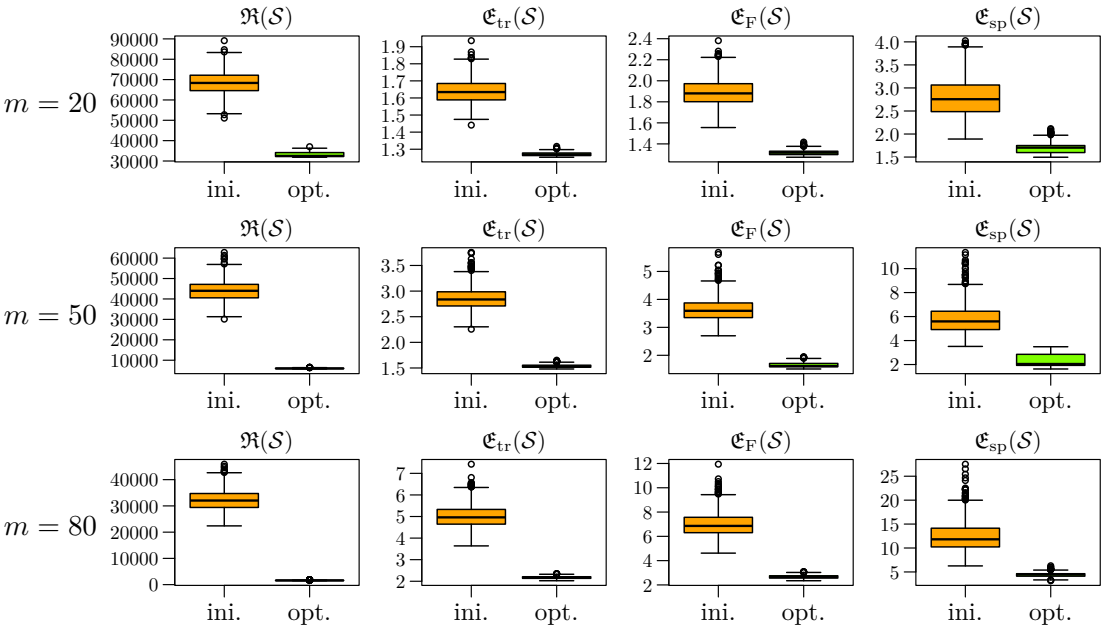

Figure 5.2: For the bi-Gaussian example, comparison of the values of $\mathfrak{R}$ and of the approximation factors (5.7) for the initial random samples $\mathcal{S}^{(0)}$ and the locally optimised samples $\mathcal{S}^{(T)}$ obtained through GD with $\rho = 10^{-6}$ and $T = 1{,}000$. Each row corresponds to a different value of the Nyström sample size m; in each case $n = 1{,}000$ repetitions are performed. The first column corresponds to $\mathfrak{R}$, and the following three correspond to the approximation factors defined in (5.7).

To further illustrate the relationship between the error map $\mathfrak{R}$ and the error maps (C.1)-(C.3), for $n = 200$ random initial samples of size $m = 15$, we perform direct minimisations, through GD, of the maps $\mathfrak{R}$ and $\mathfrak{E}_{\mathrm{tr}}$ (we consider the trace norm as it is the least costly to implement). For each descent, we assess the accuracy

of the locally-optimised Nyström samples in terms of $\mathfrak{R}$ and the trace norm; the results are presented in Figure 5.3. We observe some strong similarities between the landscapes of $\mathfrak{R}$ and $\mathfrak{C}_{\mathrm{tr}}$, further supporting the use of $\mathfrak{R}$ as a surrogate for the trace error map (the minimisation of $\mathfrak{R}$ being, from a numerical standpoint, significantly more affordable than the minimisation of the trace norm; see Section 5.2).

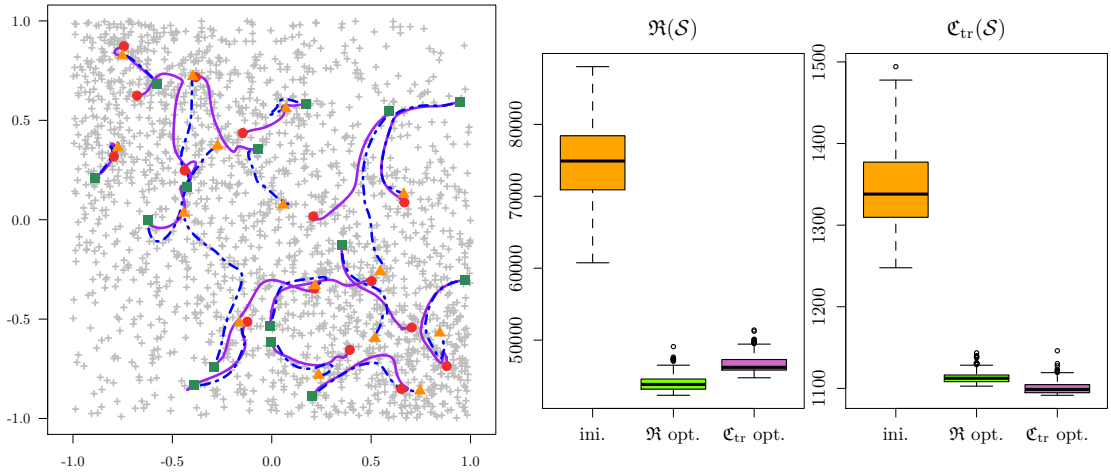

Figure 5.3: For the bi-Gaussian example of Section 5.5.1, graphical representation of the paths followed by the landmark points of a random initial Nyström sample of size $m = 15$ during the local minimisations of $\mathfrak{R}$ and $\mathfrak{C}_{\mathrm{tr}}$ through GD; the green squares are the initial landmark points, and the red dots and orange triangles are the optimised landmark points for $\mathfrak{R}$ and $\mathfrak{C}_{\mathrm{tr}}$, respectively. The solid purple lines correspond to the paths followed by the points during the minimisation of $\mathfrak{R}$, and the dashed blue lines to the paths followed during the minimisation of $\mathfrak{C}_{\mathrm{tr}}$ (left). For $n = 200$ random initial Nyström samples of size $m = 15$, comparison of the improvements yielded by the minimisations of $\mathfrak{R}$ and $\mathfrak{C}_{\mathrm{tr}}$ in terms of $\mathfrak{R}$ (middle) and trace norm (right). Each GD uses $T = 1{,}000$ iterations, with $\rho = 10^{-6}$ for $\mathfrak{R}$ and $\rho = 8 \times 10^{-5}$ for $\mathfrak{C}_{\mathrm{tr}}$.

5.5.2 Abalone data set

We now consider the $d = 8$ attributes of the Abalone data set. After removing two observations that are clear outliers, we are left with $N = 4{,}175$ entries. Each of the eight features is standardised such that it has zero mean and unit variance. We set $m = 50$ and consider three different values of the Gaussian kernel parameter γ, namely $\gamma = 0.25$, 1, and 4; these values are chosen so that the eigenvalues of the kernel matrix $\mathbf{K}$ exhibit sharp, moderate and shallower decays, respectively. For the Nyström sample optimisation, we use SGD with i.i.d. sampling and batch size $b = 50$, $T = 10{,}000$ and $\rho = 8 \times 10^{-7}$; these values were chosen to obtain relatively efficient optimisations for the whole range of values of γ we consider. For each value of γ, we perform $n = 200$ repetitions. The results are presented in Figure 5.4.

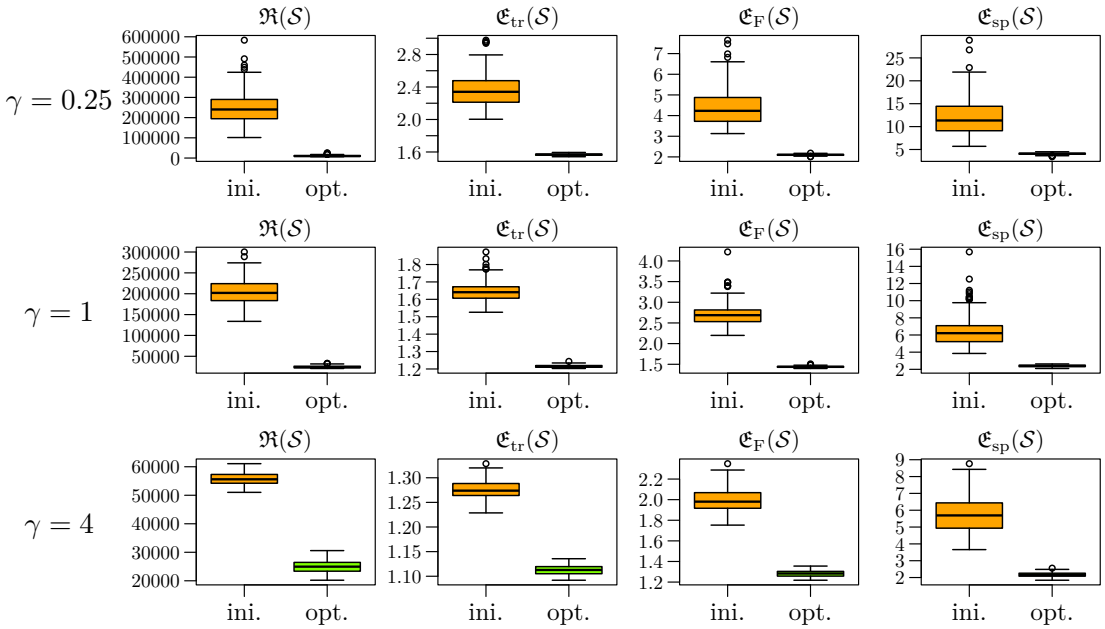

Figure 5.4: For the Abalone data set example of Section 5.5.2 with $m = 50$ and Gaussian kernel parameter $\gamma \in \{0.25, 1, 4\}$, comparison of the values of $\mathfrak{R}$ and of the approximation factors (5.7) for the initial Nyström samples $\mathcal{S}^{(0)}$ and the locally optimised samples $\mathcal{S}^{(T)}$ obtained through SGD with i.i.d. sampling ($b = 50$, $\rho = 8 \times 10^{-7}$ and $T = 10{,}000$). Each row corresponds to a given value of γ; in each case, $n = 200$ repetitions are performed.

We observe that regardless of the value of γ, in comparison with the initial Nyström samples, the accuracies of the locally optimised samples in terms of trace, Frobenius and spectral norms are significantly improved. As observed in Section 5.5.1, the gains yielded by the local optimisations are more evident in terms of trace and Frobenius norms, and the impact of the initialisation appears limited.

5.5.3 MAGIC data set

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

In Figure 5.5, we present the results obtained after the local optimisation of $n = 200$ random initial Nyström samples of size $m = 100$ and 200. Each optimisation was performed through SGD with i.i.d. sampling, batch size $b = 50$ and stepsize $\rho = 5 \times 10^{-8}$; for $m = 100$, we used $T = 3{,}000$ iterations, and $T = 4{,}000$ iterations for $m = 200$. The optimisation parameters were chosen to obtain relatively efficient but not fully completed descents, as illustrated in

Figure 5.5. Alongside the error map $\mathfrak{R}$, we only compute the approximation factor corresponding to the trace norm, since the trace norm is indeed the least costly to evaluate of the three matrix norms we consider (see Table 2.1 in Section 2.1.1). As in the previous experiments, we observe a significant improvement of the initial Nyström samples obtained by local optimisation of $\mathfrak{R}$.

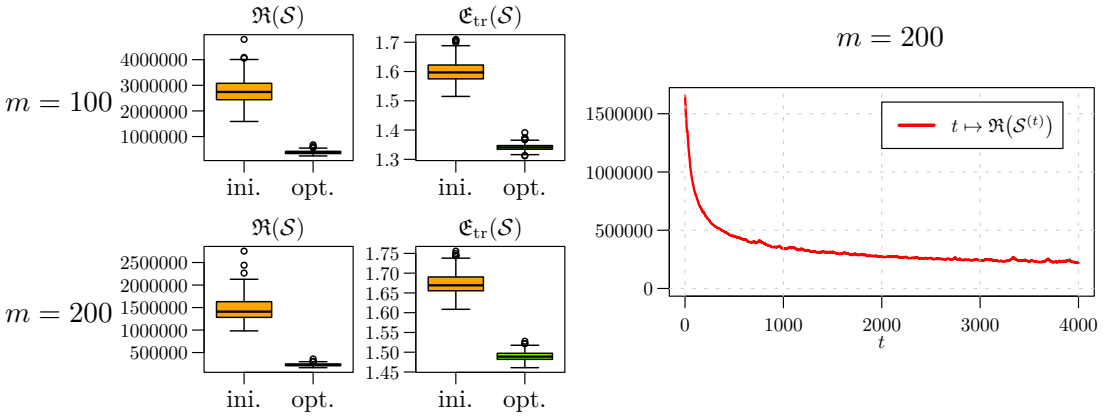

Figure 5.5: For the MAGIC data set example of Section 5.5.3, boxplots of the error map $\mathfrak{R}$ and of the approximation factor $\mathfrak{E}_{\mathrm{tr}}$ before and after the local optimisation via SGD of random Nyström samples of size $m = 100$ and 200; for each value of m, $n = 200$ repetitions are performed. The SGD is based on i.i.d. sampling, with $b = 50$ and $\rho = 5 \times 10^{-8}$; for $m = 100$, the descent is stopped after $T = 3{,}000$ iterations, and after $T = 4{,}000$ iterations for $m = 200$ (left). A graphical representation of the decay of $\mathfrak{R}$ is also presented for $m = 200$ (right).

5.5.4 MiniBooNE data set

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

We consider a random initial Nyström sample of size $m = 1{,}000$, and optimise it through SGD with i.i.d. sampling (batch size $b = 200$ and step size $\rho = 2 \times 10^{-7}$); the descent is stopped after $T = 8{,}000$ iterations. The resulting decay of the error map $\mathfrak{R}$ is presented in Figure 5.6 (the cost is evaluated every 100 iterations), and the trace norm of the Nyström approximation error for the initial and locally optimised samples are reported.

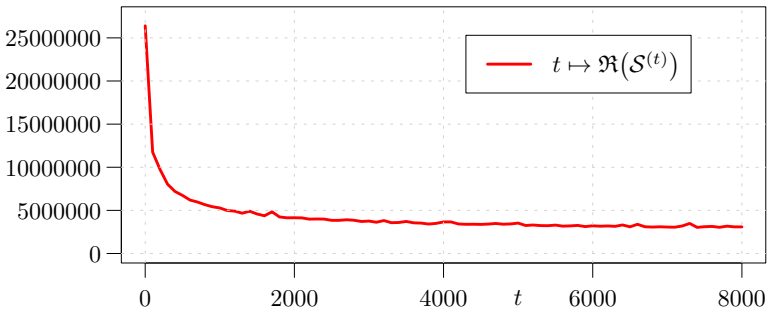

Figure 5.6: For the MiniBooNE data set of Section 5.5.4, decay of the error map $\mathfrak{R}$ during the optimisation of a random initial Nyström sample of size $m = 1{,}000$. The SGD is based on i.i.d. sampling with $b = 200$ and $\rho = 2 \times 10^{-7}$, and the descent is stopped after $T = 8{,}000$ iterations; the cost is evaluated every 100 iterations.

In terms of computational time (and for our experimental setup), in this specific example, the full optimisation process of $\mathfrak{R}$ (without checking the decay of the cost) is roughly five times faster than a single evaluation of the trace error.

Appendix: Proof of Theorem 5.1

In this appendix, we prove Theorem 5.1 of Section 5.3.

Proof of Theorem 5.1. We consider a Nyström sample $\mathcal{S} \in \mathscr{X}^m$ and introduce the quantity

$$c_{\mathcal{S}} = \frac{1}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2} \sum_{i=1}^{N} \sum_{j=1}^{m} K^2(x_i, s_j). \tag{5.8}$$

In view of (5.4), the partial derivative of $\mathfrak{R}$ at $\mathcal{S}$ with respect to the l-th coordinate of the k-th landmark point s_k can be written as

$$\partial_{[s_k]_l} \mathfrak{R}(\mathcal{S}) = c_{\mathcal{S}}^2 \left(\partial_{[s_k]_l}^{[\mathrm{d}]} K^2(s_k, s_k) + 2 \sum_{\substack{j=1, \\ j \neq k}}^{m} \partial_{[s_k]_l}^{[\mathrm{l}]} K^2(s_k, s_j) \right) - 2c_{\mathcal{S}} \sum_{i=1}^{N} \partial_{[s_k]_l}^{[\mathrm{l}]} K^2(s_k, x_i). \tag{5.9}$$

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

$$\begin{aligned} \partial_{[s_k]_l} \partial_{[s_k]_{l'}} \mathfrak{R}(\mathcal{S}) = {} & c_{\mathcal{S}}^2 \partial_{[s_k]_l}^{[\mathrm{d}]} \partial_{[s_k]_{l'}}^{[\mathrm{d}]} K^2(s_k, s_k) + 2c_{\mathcal{S}} (\partial_{[s_k]_{l'}} c_{\mathcal{S}}) \partial_{[s_k]_l}^{[\mathrm{d}]} K^2(s_k, s_k) \\ & + 2c_{\mathcal{S}}^2 \sum_{\substack{j=1, \\ j \neq k}}^{m} \partial_{[s_k]_l}^{[\mathrm{l}]} \partial_{[s_k]_{l'}}^{[\mathrm{l}]} K^2(s_k, s_j) + 4c_{\mathcal{S}} (\partial_{[s_k]_{l'}} c_{\mathcal{S}}) \sum_{\substack{j=1, \\ j \neq k}}^{m} \partial_{[s_k]_l}^{[\mathrm{l}]} K^2(s_k, s_j) \\ & - 2c_{\mathcal{S}} \sum_{i=1}^{N} \partial_{[s_k]_l}^{[\mathrm{l}]} \partial_{[s_k]_{l'}}^{[\mathrm{l}]} K^2(s_k, x_i) - 2(\partial_{[s_k]_{l'}} c_{\mathcal{S}}) \sum_{i=1}^{N} \partial_{[s_k]_l}^{[\mathrm{l}]} K^2(s_k, x_i), \end{aligned} \tag{5.10}$$

and

$$\begin{aligned}\partial_{[s_k]_l}\partial_{[s_{k'}]_{l'}}\mathfrak{R}(\mathcal{S}) = 2c_{\mathcal{S}}(\partial_{[s_{k'}]_{l'}}c_{\mathcal{S}})\partial^{[\mathrm{d}]}_{[s_k]_l}K^2(s_k,s_k) + 2c_{\mathcal{S}}^2\partial^{[\mathrm{l}]}_{[s_k]_l}\partial^{[\mathrm{r}]}_{[s_{k'}]_{l'}}K^2(s_k,s_{k'}) \\ + 4c_{\mathcal{S}}(\partial_{[s_{k'}]_{l'}}c_{\mathcal{S}})\sum_{\substack{j=1,\\ j\neq k}}^{m}\partial^{[\mathrm{l}]}_{[s_k]_l}K^2(s_k,s_j) \\ - 2(\partial_{[s_{k'}]_{l'}}c_{\mathcal{S}})\sum_{i=1}^{N}\partial^{[\mathrm{l}]}_{[s_k]_l}K^2(s_k,x_i);\end{aligned} \tag{5.11}$$

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

$$\partial_{[s_k]_l}c_{\mathcal{S}} = \frac{1}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2}\left(\sum_{i=1}^{N}\partial^{[\mathrm{l}]}_{[s_k]_l}K^2(s_k,x_i) - c_{\mathcal{S}}\partial^{[\mathrm{d}]}_{[s_k]_l}K^2(s_k,s_k) - 2c_{\mathcal{S}}\sum_{\substack{j=1,\\ j\neq k}}^{m}\partial^{[\mathrm{l}]}_{[s_k]_l}K^2(s_k,s_j)\right). \tag{5.12}$$

From (A.1), we have

$$\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}^2 = \sum_{i=1}^{m}\sum_{j=1}^{m}K^2(s_i,s_j) \geqslant \sum_{i=1}^{m}K^2(s_i,s_i) \geqslant m\alpha. \tag{5.13}$$

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

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{m}K^2(s_j,x_i) = \sum_{i=1}^{N}\sum_{j=1}^{m}\langle k_{s_j}^2 \,|\, k_{x_i}^2\rangle_{\mathcal{G}} = \left\langle \sum_{j=1}^{m}k_{s_j}^2 \,\middle|\, \sum_{i=1}^{N}k_{x_i}^2\right\rangle_{\mathcal{G}} \\ \leqslant \left\|\sum_{j=1}^{m}k_{s_j}^2\right\|_{\mathcal{G}}\left\|\sum_{i=1}^{N}k_{x_i}^2\right\|_{\mathcal{G}} = \|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}\|\mathbf{K}\|_{\mathrm{F}}.\end{aligned} \tag{5.14}$$

By combining (5.8) with inequalities (5.13) and (5.14), we obtain

$$0 \leqslant c_{\mathcal{S}} \leqslant \frac{\|\mathbf{K}\|_{\mathrm{F}}}{\|\mathbf{K}_{\mathcal{S}}\|_{\mathrm{F}}} \leqslant \frac{\|\mathbf{K}\|_{\mathrm{F}}}{\sqrt{m\alpha}} =: C_0. \tag{5.15}$$

Let $k \in [m]$ and let $l \in [d]$; from equation (5.12), and using inequalities (5.13) and (5.15) together with (A.2), we obtain

$$|\partial_{[s_k]_l}c_{\mathcal{S}}| \leqslant \frac{M_1}{m\alpha}[N + (2m-1)C_0] =: C_1. \tag{5.16}$$

In addition, let $k' \in [m] \setminus \{k\}$ and $l' \in [d]$; from equations (5.10), (5.11), (5.15) and (5.16), and conditions (A.2) and (A.3), we get

$$|\partial_{[s_k]_l}\partial_{[s_{k'}]_{l'}}\mathfrak{R}(\mathcal{S})| \leqslant C_0^2M_2 + 2C_0C_1M_1 + 2(m-1)C_0^2M_2 + 4(m-1)C_0C_1M_1$$

$$
\begin{aligned}
&+ 2C_0M_2N + 2C_1M_1N \\
&= (2m-1)C_0^2M_2 + (4m-2)C_0C_1M_1 + 2N(C_0M_2 + C_1M_1),
\end{aligned} \tag{5.17}
$$

and

$$
\begin{aligned}
|\partial_{[s_k]_l}\partial_{[s_{k'}]_{l'}}\mathfrak{R}(\mathcal{S})| &\leqslant 2C_0C_1M_1 + 2C_0^2M_2 + 4(m-1)C_0C_1M_1 + 2C_1M_1N \\
&= 2C_0^2M_2 + (4m-2)C_0C_1M_1 + 2NC_1M_1.
\end{aligned} \tag{5.18}
$$

For $k, k' \in [m]$, we denote by $\mathbf{B}^{k,k'}$ the $d \times d$ matrix with (l, l') entry given by (5.10) if $k = k'$, and by (5.11) otherwise. The Hessian $\nabla^2\mathfrak{R}(\mathcal{S})$ can then be represented as a block-matrix, that is

$$
\nabla^2\mathfrak{R}(\mathcal{S}) = \begin{bmatrix} \mathbf{B}^{1,1} & \dots & \mathbf{B}^{1,m} \\ \vdots & \ddots & \vdots \\ \mathbf{B}^{m,1} & \dots & \mathbf{B}^{m,m} \end{bmatrix} \in \mathbb{R}^{md \times md}.
$$

The d^2 entries of each of the m diagonal blocks of $\nabla^2\mathfrak{R}(\mathcal{S})$ are of the form (5.10), and the d^2 entries of each of the $m(m-1)$ off-diagonal blocks of $\nabla^2\mathfrak{R}(\mathcal{S})$ are of the form (5.11). From inequalities (5.17) and (5.18), we obtain

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

with

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

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

Chapter 6

Concluding discussion

In this chapter, we give a final overview of the main contributions of the thesis, and discuss some potential extensions of the presented work.

6.1 Summary of the contributions of the thesis

In Chapters 2 and 3, we investigated the connections between the low-rank approximation of integral operators and the Nyström approximation of PSD matrices. In this framework, we described two main classes of sampling strategies for Nyström approximation leveraging the properties of the energy setting.

Sequential column sampling. In Chapter 4, we described an energy-based 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 we described a stochastic approximation scheme 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 is then linear in N. For instance, the worst-case time-complexity of performing m iterations of the S-MFW variant of Algorithm 4.1 is $\mathcal{O}(m^2 + mN + \ell N)$, with in practice m and $\ell \ll N$, where ℓ is the row-sample size parameter for the approximation of the target potential $\mathfrak{g}$; the algorithm then extracts a sample of m columns.

We presented a series of experiments which demonstrate the ability of the proposed sequential sampling strategies to produce accurate Nyström approxi-

mations 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.

Particle-flow sampling. In Chapter 5, we described the landmark-point framework for the Nyström approximation of PSD kernel matrices. We introduced an energy-based error map $\mathfrak{R}$ which is differentiable on the landmark-point space, and provided sufficient conditions for the Lipschitz continuity of the gradient of this error map, ensuring the convergence of gradient-descent iterates with suitable step sizes. Stochastic approximations of the partial derivatives of $\mathfrak{R}$ were discussed, and we described a stochastic gradient descent procedure for the local optimisation of an initial Nyström sample of landmark points. We performed numerical experiments on a range of data sets, and observed that optimising column samples drawn uniformly at random led to consistent improvements in the quality of the induced Nyström approximations. We also demonstrated the ability of the proposed approach to handle large-scale problems.

As a side note, in addition to the two main classes of sampling strategies discussed in this thesis (that is, sequential and particle-flow-based techniques), a third type of energy-based approach to the CSP may be considered, based on sparsity-inducing regularisation; this was investigated in (Gauthier and Suykens, 2018).

6.2 Extensions and future work

In this section, we discuss potential extensions of our investigations that are directly related to the presented work. We also suggest possible directions for future research in line with the developed methodologies.

6.2.1 Extensions of the presented work

Regarding the sequential strategies described in Chapter 4, 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.

Regarding the particle-flow-based strategies described in Chapter 5, in our experiments, we used stochastic gradient descents with i.i.d. sampling, fixed step size and fixed number of iterations. Although already bringing satisfactory results, to improve the efficiency of the approach, the optimisation could be accelerated by considering for instance adaptive step sizes or momentum-type techniques (see Bottou et al., 2018 for an overview), and parallelisation may be implemented. The initial Nyström samples we considered were drawn uniformly at random without replacement; while our experiments suggest that the local minima of the error map $\mathfrak{R}$ often induce approximations of comparable quality, the use of more efficient initialisation strategies could be investigated (for instance, the sequential sampling strategies of Chapter 4 could be used to design initial samples).

6.2.2 Other research directions

The extent to which the energy setting may be adapted to the low-rank approximation of general matrices (that is, not necessarily PSD) could be an interesting avenue for future research. For instance, a general matrix $\mathbf{X}$ actually characterises two RKHSs, the RKHS $\mathcal{H}_1$ related to $\mathbf{X}^*\mathbf{X}$ and the RKHS $\mathcal{H}_2$ related to $\mathbf{X}\mathbf{X}^*$; the matrix $\mathbf{X}$ can then be regarded as an operator from $\mathcal{H}_1$ to $\mathcal{H}_2$. Further investigating the implications of this interpretation could potentially lead to some interesting developments in the CSSP setting (see Chapter 2) and for the approximation of the SVD of large matrices.

The rescaling-invariance mechanism described in Chapter 4 could also be applied to the solving of large unconstrained quadratic programs. In this setting, the introduction of suitable restrictions on the search space could benefit the development of conditional-gradient-based optimisation strategies for the minimisation of quadratic functions on high-dimensional spaces (and consequently, could support the development of novel strategies for approximating the solutions to large systems of linear equations).

More generally, the enforcement of rescaling invariances, its connections to generalised convexity and inherent suitability for the implementation of conditional-gradient-type optimisation techniques could be worth exploring in the wider context of approximate linear algebra and large-scale optimisation.

References

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson–Lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of Computing*, pages 557–563, 2006.

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

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. *arXiv preprint arXiv:1203.4523*, 2012.

João Carlos Alves Barata and Mahir Saleh Hussein. The moore-penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42:146–165, 2012.

Avrim Blum. Random projection, margins, kernels, and feature-selection. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pages 52–68. Springer, 2005.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Christos Boutsidis, Michael W Mahoney, and Petros Drineas. Column subset selection for unsupervised feature selection. 2007.

Yifan Chen and Yun Yang. Fast statistical leverage score approximation in kernel ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2935–2943. PMLR, 2021.

Yutian Chen, Max Welling, and Alexander J. Smola. Super-samples from kernel herding. In *Uncertainty in Artificial Intelligence*, 2010.

Farah Cherfaoui, Hachem Kadri, and Liva Ralaivola. Scalable ridge leverage score sampling for the Nyström method. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4163–4167. IEEE, 2022.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

Michael B Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 181–190, 2015.

Enrico Crovini, Simon L Cotter, Konstantinos Zygalakis, and Andrew B Duncan. Batch bayesian optimization via particle gradient flows. *arXiv preprint arXiv:2209.04722*, 2022.

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

Michal Derezinski, Daniele Calandriello, and Michal Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012.

Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2019. URL `http://archive.ics.uci.edu/ml`.

Nelson Dunford and Jacob T Schwartz. *Linear Operators: Spectral Theory*. Interscience Publ., 1975.

Ahmed K Farahat, Ahmed Elgohary, Ali Ghodsi, and Mohamed S Kamel. Greedy column subset selection for large-scale data sets. *Knowledge and Information Systems*, 45:1–34, 2015.

Bertrand Gauthier. Kernel embedding of measures and low-rank approximation of integral operators. *Positivity*, 2024.

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

Jean Ginibre. Statistical ensembles of complex, quaternion, and real matrices. *Journal of Mathematical Physics*, 6(3):440–449, 1965.

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

Matthew Hutchings and Bertrand Gauthier. Energy-based sequential sampling for low-rank PSD-matrix approximation. *HAL preprint hal-04102664*, 2023a.

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

Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

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

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257. PMLR, 2016.

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

Odile Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

Giacomo Meanti, Luigi Carratino, Ernesto De Vito, and Lorenzo Rosasco. Efficient hyperparameter tuning for large scale kernel ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 6554–6572. PMLR, 2022.

Madan Lal Mehta and Michel Gaudin. On the density of eigenvalues of a random matrix. *Nuclear Physics*, 18:420–427, 1960.

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

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

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

Cameron Musco and Christopher Musco. Recursive sampling for the Nyström method. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Roger Penrose. A generalized inverse for matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 406–413. Cambridge University Press, 1955.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

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

Steven Roman, S Axler, and FW Gehring. *Advanced linear algebra*, volume 3. Springer, 2005.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Shiliang Sun, Jing Zhao, and Jiang Zhu. A review of Nyström methods for large-scale machine learning. *Information Fusion*, 26:36–48, 2015.

Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29, 2016.

Lloyd N Trefethen and David Bau. *Numerical linear algebra*. SIAM, 2022.

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

Xingzhi Zhan. *Matrix inequalities*. Springer, 2004.