Random Convex Hulls and Kernel Quadrature

Satoshi Hayakawa
St Catherine's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2023

Acknowledgements

First, I would like to express my deepest gratitude to my family, Saeko, Hisao, and Hisashi, for their invaluable support throughout my journey up to this point. Without their help, I could not have pursued my curiosity about mathematics from childhood.

I also deeply thank my colleagues in Oxford. Christina, Cris, James, and Patric greatly helped me regarding life in Oxford as a maths DPhil student. I am especially grateful to Jake, my first neighbor in Oxford, for cheering me up during the lockdown, without which I could not have gone through those depressing days.

I also thank all my academic collaborators during DPhil: Ken'ichiro Tanaka, Masaki Adachi, Saad Hamid, Xingchen Wan, Martin Jørgensen, Michael Osborne, Hayata Yamasaki, Sathyawageeswar Subramanian, Sho Sonoda, and my supervisors. I have to say a special thank you to Masaki for turning my mathematical ideas into something really useful.

The university table tennis club is my best community in Oxford. I love all my teammates and especially thank Zhongyi and Haichuan for being my best friends during and not during the training. Thank you to the men's first team members: Kin for coaching and driving, Angus for being a good captain, and Alistair for that lovely two-handed backhands in our doubles. Finally, I am grateful to Taylor and Beverly for being my great rivals. I always enjoyed chopping your forehands.

I want to thank my dearest friends in Japan, including Catan-kai members. I appreciate your help and the time we spent together, without which completing this thesis would have been impossible.

I thank Toyota Riken Overseas Scholarship, Clarendon Scholarship, Oxford-Kobe Scholarship, Datasıg, and the CIMDA-Oxford for financial support on life in Oxford and academic activities.

I am deeply grateful to my examiners, Prof. Yuji Nakatsukasa and Prof. Chris Oates for assessing this thesis. Thank you to Benjamin Walker, Prof. Ken'ichiro Tanaka for proofreading the manuscript of the thesis. I am also grateful to Prof. Oliver Riordan, Prof. Raphael Hauser, and Prof. Massimiliano Gubinelli for assessing my transfer and confirmation theses. Thank you to all the anonymous reviewers of my papers for your strict comments, which improved the thesis a lot.

Last but not least, I am immensely grateful to my supervisors, Prof. Terry Lyons and Prof. Harald Oberhauser, for their help throughout my DPhil. They always patiently supported me and gave me countless ideas on life and mathematics.

Abstract

Discretization of probability measures is ubiquitous in the field of applied mathematics, from classical numerical integration to data compression and algorithmic acceleration in machine learning. In this thesis, starting from generalized Tchakaloff-type cubature, we investigate random convex hulls and kernel quadrature.

In the first two chapters after the introduction, we investigate the probability that a given vector θ is contained in the convex hull of independent copies of a random vector X. After deriving a sharp inequality that describes the relationship between the said probability and Tukey's halfspace depth, we explore the case $\theta = \mathbb{E}[X]$ by using moments of X and further the case when X enjoys some additional structure, which are of primary interest from the context of cubature.

In the subsequent two chapters, we study kernel quadrature, which is numerical integration where integrands live in a reproducing kernel Hilbert space. By explicitly exploiting the spectral properties of the associated integral operator, we derive convex kernel quadrature with theoretical guarantees described by its eigenvalue decay. We further derive practical variants of the proposed algorithm and discuss their theoretical and computational aspects.

Finally, we briefly discuss the applications and future work of the thesis, including Bayesian numerical methods, in the concluding chapter.

Publications

This thesis is based on the following original papers [71, 72, 70, 73] on the theory or application of discretization of measures. The list follows the order of appearance in the manuscript (Chapters 2, 3, 4, and 5, respectively).

[71] **Estimating the probability that a given vector is in the convex hull of a random sample**
Satoshi Hayakawa, Terry Lyons, Harald Oberhauser
Probability Theory and Related Fields, 185, 705–746, 2023

[72] **Hypercontractivity meets random convex hulls: Analysis of randomized multivariate cubatures**
Satoshi Hayakawa, Harald Oberhauser, Terry Lyons
Proceedings of the Royal Society A, 479(2273), 20220725, 2023

[70] **Positively weighted kernel quadrature via subsampling**
Satoshi Hayakawa, Harald Oberhauser, Terry Lyons
Advances in Neural Information Processing Systems, 35, 6886–6900, 2022

[73] **Sampling-based Nyström approximation and kernel quadrature**
Satoshi Hayakawa, Harald Oberhauser, Terry Lyons
Proceedings of the 40th International Conference on Machine Learning, 12678–12699, 2023

Breakdown of contributions

In all the above-mentioned publications, the authors are listed in the order of contribution and I contributed all the mathematical details, all the coding, and most of the writing. Prof. Terry Lyons posed a critical question on the behavior of $p_{n,X}$ in Hayakawa et al. [71], and supervised all four papers. Prof. Harald Oberhauser suggested relevant literature for each paper (Tukey depth and random matrices [71], hypercontractivity [72], Nyström approximation [70, 73]), helped some writing [72, 70, 73], and supervised all four papers.

Other papers

There are some more publications during my DPhil period. They are not included in this thesis; some of them were primarily conducted as my master's study at the University of Tokyo [66, 68, 69], while the others cover related but different topics such as Bayesian numerical methods [1, 2, 3] (these are briefly explained in Chapter 6) and quantum computation [180].

Contents

1 Introduction 1

1.1 Cubature and random convex hulls 1

1.2 RKHS and kernel quadrature 3

1.3 Overview and outline 4

2 Estimating the probability that a given vector is in the convex hull of a random sample 8

2.1 Introduction 8

2.1.1 Statistical depth 10

2.1.2 Inclusion of deterministic convex bodies 10

2.1.3 Organization of the chapter 12

2.2 General bounds of $p_{n,X}$ 14

2.3 Uniform bounds of $p_{n,X}^{\varepsilon}$ via the relaxed Tukey depth 18

2.4 Bounds of N_X via Berry–Esseen theorem 22

2.5 Deterministic interior body of random polytopes 24

2.6 Application 26

2.6.1 Bounds of $p_{n,X}$ 27

2.6.2 Bounds of N_X 28

2.7 Concluding remarks 31

2.A Proofs 31

2.A.1 Proof of Proposition 2.3 31

2.A.2 Proof of Proposition 2.4 32

2.A.3 Proof of Theorem 2.7 33

2.A.4 Proof of Proposition 2.8 34

2.A.5 Proof of Proposition 2.10 35

2.A.6 Proof of Lemma 2.11 35
2.A.7 Proof of Theorem 2.12 36
2.A.8 Proof of Proposition 2.14 37
2.A.9 Proof of Proposition 2.15 39
2.A.10 Proof of Theorem 2.17 40
2.A.11 Proof of Corollary 2.18 40
2.A.12 Proof of Proposition 2.19 41
2.A.13 Proof of Proposition 2.20 42
2.A.14 Proof of Proposition 2.21 42
2.A.15 Proof of Theorem 2.22 43
2.A.16 Proof of Corollary 2.23 43
2.B Bounds of N_X via multivariate Berry–Esseen theorem 44
2.B.1 Multivariate Berry–Esseen bounds 45
2.B.2 The first bound 47
2.B.3 The second bound 48
2.C Extreme examples 51

3 Hypercontractivity meets random convex hulls **56**
3.1 Introduction 56
3.2 Random convex hulls 60
3.3 Hypercontractivity 61
3.4 Applications 66
3.4.1 Classical polynomial cubatures 66
3.4.2 Cubature on Wiener space 67
3.4.3 Kernel quadrature for product measures 69
3.5 Concluding remarks 74
3.A Proofs 74
3.A.1 Proof of Proposition 3.4(b) 74
3.A.2 Proof of Corollary 3.5 75
3.A.3 Proof of Proposition 3.6 76
3.A.4 Proof of Theorem 3.15 77
3.A.5 Proof of Proposition 3.16 77
3.A.6 Proof of Proposition 3.17 78

3.A.7 Proof of Theorem 3.18 79
3.A.8 Proof of Proposition 3.25 79
3.A.9 Proof of Proposition 3.26 80

4 Positively weighted kernel quadrature via subsampling 82

4.1 Introduction 82
4.2 Main result 86
4.2.1 Algorithm 88
4.2.2 Mercer approximation 90
4.2.3 Nyström approximation 90
4.2.4 Kernel quadrature using expectations of test functions . . . 93
4.3 Numerical experiments 94
4.3.1 Periodic Sobolev spaces with uniform measure 95
4.3.2 Measure reduction in machine learning datasets 98
4.4 Concluding remarks 100
4.A Outline of Appendix 100
4.B Theoretical results and remarks 101
4.B.1 Quantitative results 101
4.B.2 Existence results 103
4.B.3 Eigenvalue estimate for Gaussian kernels 104
4.B.4 Uniform robustness 105
4.B.5 Equivalence between the projection/matrix Nyström approximations 105
4.C Proofs 106
4.C.1 Proof of Theorem 4.6 106
4.C.2 Proof of Theorem 4.7 108
4.C.3 Proof of Proposition 4.8 110
4.C.4 Proof of Theorem 4.9 110
4.C.5 Proof of Theorem 4.3 111
4.D Kernel quadrature when expectations are known 114
4.D.1 Algorithm 4.2 with Mercer approximation 117
4.D.2 Algorithm 4.2 with Nyström approximation 119
4.E Additional numerical experiments 120

4.E.1 Periodic Sobolev spaces with uniform measure 120
4.E.2 Measure reduction in machine learning datasets 121

5 Sampling-based Nyström approximation and kernel quadrature 126
5.1 Introduction 126
5.2 Related literature and notation 130
5.3 Analyzing k_s^Z for i.i.d. Z via statistical learning theory 132
5.4 A refined low-rank approximation with general Z 134
5.4.1 Part A: Estimating the error of k^Z for general Z 135
5.4.2 Part B: Mercer decomposition of k^Z 136
5.5 Application to kernel quadrature 140
5.5.1 Kernel recombination 140
5.5.2 Numerical examples 142
5.6 Concluding remarks 145
5.A Tools from statistical learning theory 146
5.B Proofs 149
5.B.1 Properties of the pseudo-inverse 149
5.B.2 Proof of Lemma 5.1 149
5.B.3 Proof of Lemma 5.2 150
5.B.4 Proof of Lemma 5.3 150
5.B.5 Proof of Theorem 5.4 151
5.B.6 Proof of Remark 5.1 156
5.B.7 Proof of Proposition 5.6 156
5.B.8 Proof of Corollary 5.7 158
5.B.9 Proof of Lemma 5.8 158
5.B.10 Proof of Lemma 5.9 158
5.B.11 Proof of Proposition 5.10 159
5.B.12 Proof of Proposition 5.13 160
5.B.13 Proof of Proposition 5.14 161
5.B.14 Proof of Theorem 5.16 163

6 Conclusion 164
6.1 Bayesian numerical methods 164
6.2 Future directions 166

Bibliography 168

List of Figures

4.1 Periodic Sobolev spaces with kernel $k_r^{\otimes d}$: The average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 20 trials is plotted for each method of obtaining Q_n. The shaded regions show their standard deviation. The worst computational time per trial was 57 seconds of **Thin + opt** in $(d, r, n) = (3, 3, 128)$, where **Thinning** was 56 seconds and **N. + emp** [+ **opt**] was 22 seconds. 96

4.2 Measure reduction in Gaussian RKHS with two ML datasets: The average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 20 trials is plotted for each method of obtaining Q_n. The shaded regions show their standard deviation. The worst computational time per trial was 14 seconds of **Thinning** [+ **opt**] in Power Plant data with $n = 128$, where **N.** + **emp** [+ **opt**] was 6.3 seconds. 99

4.3 Periodic Sobolev spaces with kernel $k_r^{\otimes d}$: The average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 50 trials is plotted for each method of obtaining Q_n. The shaded regions are sample standard deviations. The worst computational time per trial was 5 seconds of **N. + emp** and **N. + emp + opt** in $(d, r, n) = (3, 3, 80)$, while **Nyström** and **Nyström + opt** ran in 0.9 seconds under the same setting. There were no infeasible LPs. 122

4.4 Supplemental experiments for Figure 4.3. n is all odd. 123

4.5 Measure reduction in Gaussian RKHS with two ML datasets: The average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 50 trials is plotted for each method of obtaining Q_n. The shaded regions are sample standard deviations. The worst computational time per trial was 13 seconds of **N. + emp** and **N. + emp + opt** in 3D Road Network data with $n = 160$. There were 7 infeasible LPs (and 800 feasible LPs) in the experiment (a) with **Nyström** or **Nyström + opt**. There were no infeasible LPs in (b). 123

4.6 Measure reduction in Gaussian RKHS with two ML datasets with another empirical measure: The average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu')^2)$ over 20 trials is plotted for each method of obtaining Q_n. The shaded regions show their standard deviation. The worst computational time per trial was 14 seconds of **Thinning** [+ **opt**] in Power Plant data with $n = 128$, where **N. + emp** [+ **opt**] was 6.2 [6.1] seconds. 125

5.1 Experiments in periodic Sobolev spaces with reproducing kernel $k_r^{\otimes d}$. Average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 20 samples plotted with their standard deviation. 143

5.2 Experiments in k_2 with $N = n^2, n^3$ for recombination algorithms. Average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 20 samples plotted with their standard deviation. 144

Chapter 1

Introduction

In this thesis, we tackle the problem of discretizing probability measures. This involves numerically approximating integrals, finding efficient samples to represent randomness in nature or big data, and accelerating scientific computing that involves randomness. Although the concrete topics of the thesis cover random convex hulls and kernel quadrature, they both originate from the following concept of discretization — *cubature.*

1.1 Cubature and random convex hulls

Let μ be a Borel probability measure on some topological space $\mathcal{X}$. Consider d integrable functions $\varphi_1, \dots, \varphi_d : \mathcal{X} \to \mathbb{R}$. We denote $\boldsymbol{\varphi} := (\varphi_1, \dots, \varphi_d)^\top$, the d-dimensional vector-valued function. Then, we know there exists a "good reduction" of μ with respect to $\boldsymbol{\varphi}$ by the following theorem:

Theorem 1.1. *There are n points $x_1, \dots, x_n \in \operatorname{supp}\mu$ and weights $w_1, \dots, w_n \geq 0$ with $n \leq d+1$ such that $w_1 + \cdots + w_n = 1$ and*

$$\int_{\mathcal{X}} \boldsymbol{\varphi}(x) \, \mathrm{d}\mu(x) = \sum_{i=1}^{n} w_i \boldsymbol{\varphi}(x_i). \tag{1.1}$$

This is often referred to as Tchakaloff's theorem [162, 12], although a more accurate nomenclature would involve Wald, Richter, Rogosinski, and Rosenbloom [175, 162, 141, 142, 143]; see di Dio and Schmüdgen [40] for a historical perspective.

The proof is essentially given by classical Carathéodory's theorem [25]. The points and weights treated in Tchakaloff's theorem are called cubature [161] — an important object in the field of numerical integration. An equivalent problem is also treated as a useful way to compress data in the field of data science [111, 34]. When $\mathcal{X}$ is a subset of a Euclidean space, monomials are a typical choice of test functions φ_i, so the the discrete measure $\sum_{i=1}^n w_i \delta_{x_i}$ is a good approximation of μ in terms of smooth integrands. However, constructions in general settings are also useful; for example, in the cubature on Wiener space [110], $\mathcal{X}$ is the space of continuous paths, μ is the Wiener measure, and the test functions are iterated integrals of paths.

To the generalized cubature construction (or measure reduction) problem, when μ is discrete, there are efficient deterministic approaches [103, 163, 111]. In the general case, although there is no universal efficient algorithm, we can consider the following naive sampling-based approach. We are given a d-dimensional random vector $X = \boldsymbol{\varphi}(Y)$. Given its independent copies $X_i = \boldsymbol{\varphi}(Y_i)$ for $i = 1, 2, \dots, n$, once $\mathbb{E}[X] \in \operatorname{conv}\{X_1, \dots, X_n\}$ is satisfied, by using a linear programming (LP) solver, we can find sparse weights $w_i \geq 0$ with $\sum_{i=1}^n w_i = 1$ such that

$$\mathbb{E}[\boldsymbol{\varphi}(Y)] = \mathbb{E}[X] = \sum_{i=1}^n w_i X_i = \sum_{i=1}^n w_i \boldsymbol{\varphi}(Y_i),$$

which yields a generalized cubature formula. This approach is explicitly proposed by Hayakawa [66] and called *Monte Carlo cubature construction*, and it empirically works for constructing classical polynomial cubature [66] and cubature on Wiener space [68] of moderate size with a "reasonable" magnitude of n (e.g., $3d$). But there was no theoretical explanation of why it works except for the case that the distribution of X has a strong symmetry [178, 172].

Contributions on random convex hulls. To estimate the necessary computational time of this approach, we investigate the quantities

$$p_{n,X}(\theta) := \mathbb{P}(\theta \in \operatorname{conv}\{X_1, \dots, X_n\}), \quad N_X(\theta) := \inf\{n \mid p_{n,X}(\theta) \geq 1/2\},$$

for a d-dimensional random vector X and $\theta \in \mathbb{R}^d$, which are about the number of i.i.d. vectors we should sample before we get θ in the random convex hull. A sharp upper bound of $N_X(\mathbb{E}[X])$ can be used as a threshold for the sample size we need to generate in the Monte Carlo cubature construction. In the first part of this thesis, we derive sharp upper bounds of $p_{n,X}$ and lower bounds of N_X (Chapter 2) for a general random vector X as well as with a structured one $X = \boldsymbol{\varphi}(Y)$ (Chapter 3), while the previous studies have been on the bounds of the opposite directions [178, 172] or on a specific class of random vectors such as Gaussian [81].

1.2 RKHS and kernel quadrature

The setting in Section 1.1 is quite general — we have a probability measure that we want to approximate and a family of test functions. To make it usable, we need to specify the test functions or the function space of our interest; we explore kernel quadrature as a somewhat concrete objective.

On a topological space $\mathcal{X}$, we are given a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. It is called a *positive definite kernel* if the (Gram) matrix $(k(x_i, x_j))_{i,j=1}^n$ is positive semi-definite for each $x_1, \dots, x_n \in \mathcal{X}$. Then, it has an associated Hilbert space $\mathcal{H}_k$ called *reproducing kernel Hilbert space* (RKHS) with the following properties:

- $\operatorname{span}\{k(\cdot, x) \mid x \in \mathcal{X}\}$ is a dense subspace of $\mathcal{H}_k$;
- $\langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} = f(x)$ holds for each $x \in \mathcal{X}$ and $f \in \mathcal{H}_k$.

See, e.g., Berlinet and Thomas-Agnan [18] for a formal introduction to RKHSs. We treat real RKHSs throughout the thesis. For an $f : \mathcal{X} \to \mathbb{R}$, $f \in \mathcal{H}_k$ if and only if there is a constant $c > 0$ such that $k(x, y) - cf(x)f(y)$ is still positive definite (see Paulsen and Raghupathi [135, Theorem 3.11]; though they work with a complex RKHS, the same proof works for a real RKHS).

Let us then explain kernel quadrature. Given a Borel probability measure μ on $\mathcal{X}$, our objective is to find a good n-point *quadrature* Q_n composed of points $x_1, \dots, x_n \in \mathcal{X}$ and weights $w_1, \dots, w_n \in \mathbb{R}$ with a small worst-case error:

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu) := \sup_{\|f\|_{\mathcal{H}_k} \le 1} \left| \int_{\mathcal{X}} f(x) \,\mathrm{d}\mu(x) - \sum_{i=1}^n w_i f(x_i) \right|.$$

It gives a criterion to evaluate how well the discrete measure $\sum_{i=1}^n w_i \delta_{x_i}$ (also denoted by Q_n) approximates the target measure μ. This worst-case error is also known as the *maximum mean discrepancy* (or MMD distance) In general, given two Borel (probability) measures μ and ν with $x \mapsto \sqrt{k(x,x)} \in L^1(\mu) \cap L^1(\nu)$ (this condition can be weakened to $k(x,y) \in L^1(\mu \times \mu) \cap L^1(\nu \times \nu)$), their MMD distance $\mathrm{MMD}_k(\mu,\nu)$ is defined and can be computed as

$$\mathrm{MMD}_k(\mu,\nu) := \sup_{\|f\|_{\mathcal{H}_k} \le 1} \left| \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) - \int_{\mathcal{X}} f(x)\,\mathrm{d}\nu(x) \right|,$$

$$\begin{aligned}\mathrm{MMD}_k(\mu,\nu)^2 = &\iint_{\mathcal{X}\times\mathcal{X}} k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) - 2\iint_{\mathcal{X}\times\mathcal{X}} k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y) \\ &+ \iint_{\mathcal{X}\times\mathcal{X}} k(x,y)\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(y). \end{aligned} \tag{1.2}$$

The latter is a well-known formula for computing the MMD distance [58, 159].

Contributions on kernel quadrature. It is known that, under mild assumptions (e.g., $\operatorname{supp}\mu = \mathcal{X}$, k is continuous, and $x \mapsto k(x,x) \in L^1(\mu)$), we can have the following Mercer decomposition [160]: $k(x,y) = \sum_{i=1}^\infty \sigma_i e_i(x) e_i(y)$, where σ_i is an eigenvalue of the integral operator $\mathcal{K} : L^2(\mu) \to L^2(\mu)$ given by

$$\mathcal{K}f = \int_{\mathcal{X}} k(\cdot, x) f(x)\,\mathrm{d}\mu(x) \tag{1.3}$$

corresponding to the eigenfunction e_i, and they are ordered ($\sigma_1 \ge \sigma_2 \ge \cdots \ge 0$) and normalized ($\|e_i\|_{L^2(\mu)} = 1$). The eigenvalue decay has been known to be closely related to the worst-case error of a good n-point kernel quadrature, such as $\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \sim \sigma_n$ [6, 16, 17], but there were no practical algorithms that exploit this eigenvalue decay and are applicable to general pair of (k, μ) with access to an i.i.d. sample from μ; a detailed literature review will be given in Chapter 4 (see also Table 4.1 for a comparison of relevant methods). In Chapters 4 & 5, we shall give a novel kernel quadrature method that has, among other benefits, (1) a practical algorithm and (2) theoretical guarantees based on the above spectral decay, which have not been satisfied at the same time in the previous studies.

1.3 Overview and outline

In this thesis, starting from the generalized cubature problem, we address the following questions in Chapters 2–5:

- *Why does the naive randomized cubature construction work?*
- *How can we apply the idea to the kernel quadrature problem?*

We also briefly discuss applications and future directions of the thesis in Chapter 6.

We shall explain more details of our contributions together with the outline of the main body (Chapters 2–5) in the following.

Why — **Random convex hulls (Chapters 2 & 3).** The first two chapters after this introduction are devoted to analyzing $p_{n,X}$ and N_X introduced in Section 1.1.

In Chapter 2, we treat a general random vector X and point out that the Tukey depth [169] (or halfspace depth) defined as

$$\alpha_X(\theta) := \inf_{c \in \mathbb{R}^d \setminus \{0\}} \mathbb{P}\bigl(c^\top (X - \theta) \le 0\bigr)$$

plays an essential role in analyzing random convex hulls. Indeed, in Theorem 2.13, we prove the inequality

$$1/2 \le \alpha_X(\theta) N_X(\theta) \le 3d + 1$$

for a general X, which is sharp up to a constant. This main result further yields a bound of $N_X(\mathbb{E}[X])$ based on the moments of X (Section 2.4) as well as a description of the deterministic body included in a random convex hull with high probability (Section 2.5).

In Chapter 3, we consider the case where X is given by a vector-valued function with some structure $X = \boldsymbol{\varphi}(Y)$, e.g., $\boldsymbol{\varphi}$ is given by multivariate polynomials up to some degree. By generalizing the hypercontractivity arguments in Gaussian Wiener chaos [79, Chapter 5] and applying it to our result in Chapter 2, we prove the following result (Corollary 3.20):

> Let ℓ, m be positive integers and Z be an $\mathbb{R}$-valued random variable with $\mathbb{E}[|Z|^{4m}] < \infty$. If a d-dimensional random vector is given by $X = \boldsymbol{\varphi}(Z_1, \dots, Z_\ell)$, where $Z_1, \dots, Z_\ell$ are independent copies of Z and each coordinate of $\boldsymbol{\varphi} : \mathbb{R}^k \to \mathbb{R}^d$ is given by a polynomial up to degree m, then there is a constant $C_m > 0$ independent of ℓ such that
>
> $$N_X(\mathbb{E}[X]) \le C_m d.$$

This partially explains what we observe in Hayakawa [66], but is just an instance of our more general argument on cubature problems with product structure, and our examples also include the Monte Carlo approach to kernel quadrature (see the next section) and cubature on Wiener space [110].

How — **Kernel quadrature (Chapters 4 & 5).** The final two chapters of the thesis address the kernel quadrature problem from the viewpoint of generalized cubature. Recall we are given an RKHS $\mathcal{H}_k$ with a Borel probability measure μ, which admits the Mercer decomposition $k(x, y) = \sum_{i=1}^\infty \sigma_i e_i(x)e_i(y)$.

In Chapter 4, the essential idea behind the general theory is the use of $\boldsymbol{\varphi} = (e_1, \dots, e_{n-1})^\top$ in the context of generalized cubature (1.1). Indeed, with this $\boldsymbol{\varphi}$, we can prove that, if a *convex* quadrature $Q_n = (w_i, x_i)_{i=1}^n$ ("convex" means that the weights satisfy $w_i \ge 0$ and $\sum_{i=1}^n w_i = 1$) satisfies

$$\sum_{i=1}^n w_i \boldsymbol{\varphi}(x_i) = \int_{\mathcal{X}} \boldsymbol{\varphi}(x) \,\mathrm{d}\mu(x), \quad \sum_{i=1}^n w_i \sum_{m=n}^\infty \sigma_m e_m(x_i)^2 \le \int_{\mathcal{X}} \sum_{m=n}^\infty \sigma_m e_m(x)^2 \,\mathrm{d}\mu(x),$$

then we have $\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \le 4 \sum_{m=n}^\infty \sigma_m$ (Theorem 4.5). Indeed, such a convex quadrature can be obtained via random sampling in the spirit of the Monte Carlo cubature construction.

This construction gives a nice convergence guarantee as well as the convexity condition of weights, which makes Q_n a probability measure, but requires the knowledge of the Mercer decomposition, which is not necessarily available in a general situation. It is also beneficial to have the option of avoiding the use of random convex hulls as there are still open questions on them, despite the progress in Chapters 2 & 3. Thus, we generalize the above approach in the following two ways and confirm their performance in numerical experiments.

(a) **Use of any finite-rank approximation of $\boldsymbol{k}$.** We develop a theory that is applicable to any finite-rank kernel $k_0(x, y) = \sum_{i=1}^{n-1} c_i \varphi_i(x) \varphi_i(y)$ with $c_i \ge 0$ and $k_1 := k - k_0$ being positive definite as well. If we use $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_{n-1})^\top$ and replace the constraint by

$$\sum_{i=1}^n w_i \boldsymbol{\varphi}(x_i) = \int_{\mathcal{X}} \boldsymbol{\varphi}(x) \,\mathrm{d}\mu(x), \quad \sum_{i=1}^n w_i k_1(x_i, x_i) \le \int_{\mathcal{X}} k_1(x, x) \,\mathrm{d}\mu(x), \quad (1.4)$$

then we have $\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \le 4 \int_{\mathcal{X}} k_1(x, x) \,\mathrm{d}\mu(x)$ (Theorem 4.5), where the above method of using eigenfunctions is just a special example with $k_0(x, y) = \sum_{i=1}^{n-1} \sigma_i e_i(x) e_i(y)$.

(b) **Use of empirical measure and the recombination algorithm.** Finding a set of points and weights with (1.4) would require the use of random convex hulls, but we can consider replacing μ in (1.4) by its empirical measure $\widetilde{\mu}_N = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$ where y_i are independent samples from μ. Then, we can find such a convex kernel quadrature in $\mathcal{O}(nN + n^3 \log(N/n))$ computational steps by using the recombination algorithm [103, 163], and the resulting Q_n satisfies (Theorem 4.1)

$$\mathbb{E}\big[\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2\big] \le 8 \int_{\mathcal{X}} k_1(x, x) \,\mathrm{d}\mu(x) + \frac{2}{N} \int_{\mathcal{X}} k(x, x) \,\mathrm{d}\mu(x),$$

which yields a practical instance of our convex kernel quadrature.

Among possible choices of k_0 introduced in (a), arguably the most promising choice is the Nyström approximation [179, 43, 97]; the s-rank Nyström approximation based on ℓ ($\ge s$) landmark points $Z = (z_i)_{i=1}^{\ell} \subset \mathcal{X}$ is given by

$$k(x, y) \approx k_s^Z(x, y) := k(x, Z) k(Z, Z)_s^+ k(Z, y),$$

where $k(A, B)$ for $A = (a_i)$ and $B = (b_j)$ generally represents the matrix $(k(a_i, b_j))_{ij}$, and $k(Z, Z)_s^+$ is the Moore–Penrose pseudo-inverse of the best s-rank approximation of the Gram matrix $k(Z, Z) = (k(z_i, z_j))_{i,j=1}^{\ell}$.

In Chapter 5, we first investigate how well k_s^Z approximates the original kernel k by estimating the quantity

$$\int_{\mathcal{X}} \sqrt{k(x, x) - k_s^Z(x, x)} \,\mathrm{d}\mu(x),$$

which yields a subsequent theoretical guarantee of a relevant kernel quadrature rule, in the case where Z is given by independent samples from μ. Our analysis is based on the observation that k_s^Z is given by a truncated Mercer decomposition of the kernel $k^Z = k_\ell^Z$, and an application of statistical learning theory. By generalizing this observation, we further propose other low-rank approximations $k_{s,\mu}^Z$ and $k_{s,X}^Z$ for a non-i.i.d. Z and prove their favorable theoretical properties, which are confirmed both in theory and numerical experiments.

Chapter 2

Estimating the probability that a given vector is in the convex hull of a random sample

For a d-dimensional random vector X, let $p_{n,X}(\theta)$ be the probability that the convex hull of n independent copies of X contains a given point θ. We also define $N_X(\theta)$ as the smallest n for which $p_{n,X}(\theta) \geq 1/2$, as introduced in Section 1.1. In this chapter, we provide several sharp inequalities regarding $p_{n,X}$ and N_X. As a main result, we derive a totally general inequality, $1/2 \leq \alpha_X(\theta)N_X(\theta) \leq 3d+1$, where $\alpha_X(\theta)$ (known as the Tukey depth) is the minimum probability that X is in a fixed closed halfspace containing the point θ. We also show several applications of our general results. One is a moment-based bound on $N_X(\mathbb{E}[X])$, which is an important quantity in randomized approaches to cubature construction or measure reduction problems. Another application is the determination of the canonical convex body included in a random convex polytope given by independent copies of X, where our combinatorial approach allows us to generalize existing results in the random matrix community significantly.

2.1 Introduction

Consider generating independent and identically distributed d-dimensional random vectors. How many vectors do we have to generate in order that a point $\theta \in \mathbb{R}^d$ is contained in the convex hull of the sample with probability at least $1/2$? More generally, what is the probability of the event with an n-point sample for each n? These questions were first solved for a general distribution which has a certain symmetry about θ by Wendel [178]. Let us describe the problem more formally.

Let X be a d-dimensional random vector and $X_1, X_2, \ldots$ be independent copies of X. For each $\theta \in \mathbb{R}^d$ and positive integer n, recall we have defined

$$p_{n,X}(\theta) := \mathbb{P}(\theta \in \operatorname{conv}\{X_1, \ldots, X_n\}), \quad N_X(\theta) := \inf\{n \mid p_{n,X}(\theta) \geq 1/2\}$$

as quantities on a reasonable sample size we need (in Section 1.1). As $p_{n,X}$ and N_X are only dependent on the probability distribution of X, we also write $p_{n,\mu}$ and N_μ when X follows the distribution μ. We want to evaluate $p_{n,X}$ as well as N_X for a general X.

Wendel [178] showed that

$$p_{n,X}(0) = 1 - \frac{1}{2^{n-1}} \sum_{i=0}^{d-1} \binom{n-1}{i} \tag{2.1}$$

holds for an X such that X and $-X$ have the same distribution and $X_1, \ldots, X_d$ are almost surely linearly independent. In particular, $N_X(0) = 2d$ holds for such random vectors. For an X with an absolutely continuous distribution with respect to the Lebesgue measure, Wagner and Welzl [172] showed more generally that the right-hand side of (2.1) is indeed an upper bound of $p_{n,X}(0)$, and they also characterized the condition for equality (see Theorem 2.5). Moreover, Kabluchko and Zaporozhets [81] recently gave an explicit formula for $p_{n,X}$ when X is a shifted Gaussian.

In this chapter, our aim is to give generic bounds of $p_{n,X}$ and N_X, and we are particularly interested in the upper bound of N_X, which is opposite to the bound given by Wagner and Welzl [172]. Estimating $p_{n,X}$ and N_X is of great interest from application, which ranges from numerical analysis to statistics, and compressed sensing. As a by-product, we also give a general result explaining the deterministic body included in the random polytope $\operatorname{conv}\{X_1, \ldots, X_n\}$, which is a generalization of recent work in the random matrix community [60]. The remainder of this section will explain more detailed motivation from related fields and implications of our results.

Throughout the chapter, let $\langle \cdot, \cdot \rangle$ be any inner product on $\mathbb{R}^d$, and $\| \cdot \|$ be the norm it induces.

2.1.1 Statistical depth

From the statistical context, $p_{d+1,X}(\theta)$ for a d-dimensional X is called the simplicial depth of $\theta \in \mathbb{R}^d$ with respect to the (population) distribution of X [107, 27], which can be used for mathematically characterizing the intuitive "depth" of each point θ when we are given the distribution of X. For an empirical measure, it corresponds to the number of simplices (whose vertices are in the data) containing θ.

There are also various concepts measuring depth, all called statistical depth [27, 122]. One of the first such concepts is the halfspace depth proposed by Tukey [169]:

$$\alpha_X(\theta) := \inf_{c \in \mathbb{R}^d \setminus \{0\}} \mathbb{P}(\langle c, X - \theta \rangle \leq 0),$$

which can equivalently be defined as the minimum measure of a halfspace containing θ. Donoho and Gasko [42] and Rousseeuw and Ruts [144] extensively studied general features of α_X. We call it the Tukey depth throughout the chapter.

Our finding is that these two depth notions are indeed deeply related. We prove the rate of convergence $p_{n,X} \to 1$ is essentially determined by α_X (Proposition 2.24), and we have a beautiful relation $1/2 \leq \alpha_X N_X \leq 3d + 1$ in Theorem 2.13.

2.1.2 Inclusion of deterministic convex bodies

Although we have seen the background of the $p_{n,X}(\theta)$, which only describes the probability of a single vector contained in the random convex polytope, several aspects of such random polytopes have been studied [112, 77]. In particular, people also studied deterministic convex bodies associated with the distribution of a random vector. For example, one consequence of the well-known Dvoretzky–Milman's Theorem (see, e.g., Vershynin [171, Chapter 11]) is that the convex hull of n independent samples from the d-dimensional standard normal distribution is "approximately" a Euclidean ball of radius $\sim \sqrt{\log n}$ with high probability for a sufficiently large n.

Mainly from the context of random matrices, there have been several pieces of research on the interior convex body of $\mathrm{conv}\{X_1, \ldots, X_n\}$ or its "absolute" version $\mathrm{conv}\{\pm X_1, \ldots, \pm X_n\}$ for various classes of X such as Gaussian, Rademacher or vector with i.i.d. subgaussian entries [56, 52, 104, 37, 61]. One result on the Rademacher vector is the following:

Theorem 2.1 ([52]). *Let d be a sufficiently large positive integer and $X_1, X_2, \ldots$ be independent samples from the uniform distribution over the set $\{-1,1\}^d \subset \mathbb{R}^d$. Then, there exists an absolute constant $c > 0$ such that, for each integer $n \geq d(\log d)^2$, we have*

$$\operatorname{conv}\{\pm X_1, \ldots, \pm X_n\} \supset c\left(\sqrt{\log(n/d)}B_2^d \cap B_\infty^d\right)$$

with probability at least $1 - e^{-d}$. Here, B_2^d is the Euclidean unit ball in $\mathbb{R}^d$ and $B_\infty^d = [-1,1]^d$.

Although each of those results in literature was based on its specific assumptions on the distribution of X, Guédon et al. [60] found a possible way of treating the results in a unified manner under some technical assumptions on X. They introduced the floating body associated with X

$$\widetilde{K}^\alpha(X) := \{s \in \mathbb{R}^d \mid \mathbb{P}(\langle s, X\rangle \geq 1) \leq \alpha\}$$

to our context (the notation here is slightly changed from the original one), and argued that, under some assumptions on X, with high probability, $\operatorname{conv}\{X_1, \ldots, X_n\}$ includes a constant multiple of the polar body of $\widetilde{K}^\alpha(X)$ with $\log(1/\alpha) \sim 1 + \log(n/d)$. Note that their main object of interest is the absolute convex hull, but their results can be extended to the ordinary convex hull (see Guédon et al. [60, Remark 1.7]).

Let us explain more formally. Firstly, for a set $A \subset \mathbb{R}^d$, the polar body of A is defined as

$$A^\circ := \{x \in \mathbb{R}^d \mid \langle a, x\rangle \leq 1 \text{ for all } a \in A\}.$$

Secondly, we shall describe the assumptions used in Guédon et al. [60]. Let $|||\cdot|||$ be a norm on $\mathbb{R}^d$ and $\gamma, \delta, r, R > 0$ be constants. Their assumptions are as follows:

- (γ, δ) small-ball condition: $\mathbb{P}(|\langle t, X\rangle| \geq \gamma|||t|||) \geq \delta$ holds for all $t \in \mathbb{R}^d$.
- L_r condition with constant R: $\mathbb{E}[|\langle t, X\rangle|^r]^{1/r} \leq R|||t|||$ holds for all $t \in \mathbb{R}^d$.

Under these conditions, they proved the following assertion by using concentration inequalities.

Theorem 2.2 ([60])**.** *Let X be a d-dimensional symmetric random vector that satisfies the small-ball condition and L_r condition for a norm $\left\vert\!\left\vert\!\left\vert\cdot\right\vert\!\right\vert\!\right\vert$ and constants $\gamma, \delta, r, R > 0$. Let $\beta \in (0,1)$ and set $\alpha = (en/d)^{-\beta}$. Then, there exist a constant $c_0 = c_0(\beta, \delta, r, R/\gamma)$ and an absolute constant $c_1 > 0$ such that, for each integer $n \geq c_0 d$,*

$$\mathrm{conv}\{X_1, \ldots, X_n\} \supset \frac{1}{2}\big(\widetilde{K}^\alpha(X)\big)^\circ$$

holds with probability at least $1 - 2\exp(-c_1 n^{1-\beta} d^\beta)$, where $X_1, X_2, \ldots$ are independent copies of X.

Though computing $\big(\widetilde{K}^\alpha(X)\big)^\circ$ for individual X is not necessarily an easy task, this gives us a unified understanding of existing results in terms of the polar of the floating body $\widetilde{K}^\alpha(X)$. However, its use is limited due to technical assumptions. In this chapter, we show that we can completely remove the assumptions in Theorem 2.2 and obtain a similar statement only with explicit constants (see Proposition 2.20 and Corollary 2.23, or the next section).

Finally, we add that this interior body of random polytopes or its radius is recently reported to be essential in the robustness of sparse recovery [60] and the convergence rate of greedy approximation algorithms [119, 32] when the data is random.

2.1.3 Organization of the chapter

In this chapter, our aim is to derive general inequalities for $p_{n,X}$ and N_X. The main part of this chapter is Section 2.2 to 2.5. We first give general bounds of $p_{n,X}$ without specific quantitative assumptions in Section 2.2, and present novel bounds of $p_{n,X}$ uniformly determined by the Tukey depth α_X in Section 2.3, which is the primary contribution of this chapter. Section 2.4 then gives uniform bounds of $N_X(\mathbb{E}[X])$ based on the moments of X, while the results on deterministic convex bodies included in random polytopes are given in Section 2.5.

Let us give more detailed explanations of each section. Section 2.2 provides generalization of the results of Wagner and Welzl [172], and we give generic bounds of $p_{n,X}(\theta)$ under a mild assumption $p_{d,X}(\theta) = 0$, which is satisfied with absolutely continuous distributions as well as typical empirical distributions. Our main result in Section 2.2 is as follows (Theorem 2.7):

Theorem. *Let X be an arbitrary d-dimensional random vector and $\theta \in \mathbb{R}^d$. If $p_{d,X}(\theta) = 0$ holds, then, for any $n \geq m \geq d+1$, inequalities*

$$p_{n,X}(\theta) \leq 1 - \frac{1}{2^{n-1}} \sum_{i=0}^{d-1} \binom{n-1}{i}, \qquad \frac{1}{2^{n-m}} \frac{\binom{n}{d+1}}{\binom{m}{d+1}} p_{m,X}(\theta) \leq p_{n,X}(\theta) \leq \frac{\binom{n}{d+1}}{\binom{m}{d+1}} p_{m,X}(\theta)$$

hold.

In Section 2.3, we introduce $p^{\varepsilon}_{n,X}$ and α^{ε}_X for an $\varepsilon \geq 0$, which are "ε-relaxation" of $p_{n,X}$ and α_X in that $p^0_{n,X} = p_{n,X}$ and $\alpha^0_X = \alpha_X$ hold. For this generalization, we prove that the convergence of $p^{\varepsilon}_{n,X} \to 1$ is uniformly evaluated in terms of α^{ε}_X (Proposition 2.24), and obtain the following result (Theorem 2.12):

Theorem. *Let X be an arbitrary d-dimensional random vector and $\theta \in \mathbb{R}^d$. Then, for each $\varepsilon \geq 0$ and positive integer $n \geq 3d/\alpha^{\varepsilon}_X(\theta)$, we have*

$$p^{\varepsilon}_{n,X}(\theta) > 1 - \frac{1}{2^d}.$$

Although we do not define ε-relaxation version here, we can see from the case $\varepsilon = 0$ that, for example, $N_X(\theta) \leq \lceil 3d/\alpha_X(\theta) \rceil$ generally holds (see also Theorem 2.13).

In Section 2.4, we derive upper bounds of N_X without relying on α_X, which may also be unfamiliar. By using the result in the preceding section and the Berry–Esseen theorem, we show some upper bounds of N_X in terms of the (normalized) moments of X as follows (Theorem 2.17):

Theorem. *Let X be a centered d-dimensional random vector with nonsingular covariance matrix V. Then,*

$$N_X \leq 17d \left(1 + \frac{9}{4} \sup_{c \in \mathbb{R}^d, \|c\|_2 = 1} \mathbb{E}\left[\left| c^\top V^{-1/2} X \right|^3 \right]^2 \right)$$

holds.

Here, $\| \cdot \|_2$ denotes the usual Euclidean norm on $\mathbb{R}^d$. Note that the right-hand side can easily be replaced by the moment of $\|V^{-1/2}X\|_2$ (see also Corollary 2.18).

Section 2.5 asserts that $K^{\alpha}(X) := \{\theta \in \mathbb{R}^d \mid \alpha_X(\theta) \geq \alpha\}$ ($\alpha \in (0,1)$) is a canonical deterministic body included in the random convex polytope $\mathrm{conv}\{X_1, \ldots, X_n\}$. We see in Proposition 2.20 that this body is essentially equivalent to the $\big(\widetilde{K}^{\alpha}(X)\big)^{\circ}$ mentioned in Section 2.1.2, and prove the following (Theorem 2.22):

Theorem. *Let X be an arbitrary symmetric d-dimensional random vector, and let $\alpha, \delta, \varepsilon \in (0,1)$. If a positive integer n satisfies*

$$n \geq \frac{2d}{\alpha} \max\left\{ \frac{\log(1/\delta)}{d} + \log\frac{1}{\varepsilon},\ 6 \right\},$$

then we have, with probability at least $1-\delta$,

$$\mathrm{conv}\{X_1, \ldots, X_n\} \supset (1-\varepsilon)K^\alpha(X),$$

where $X_1, X_2, \ldots$ are independent copies of X.

A consequence of this theorem (Corollary 2.23) enables us to remove the technical assumption of Theorem 2.2.

Note that all these results give explicit constants with reasonable magnitude, which is because of our combinatorial approach typically seen in the proof of Proposition 2.9 and Proposition 2.14. After these main sections, we give some implications of our results on motivational examples in Section 2.6, and we finally give our conclusion in Section 2.7.

2.2 General bounds of $p_{n,X}$

In this section, we denote $p_{n,X}(0)$ by only $p_{n,X}$. As we always have $p_{n,X}(\theta) = p_{n,X-\theta}(0)$, it suffices to treat $p_{n,X}(0)$ unless we consider properties of $p_{n,X}$ as a function.

Let us start with easier observations. Proposition 2.3 and Proposition 2.4 are almost dimension-free. Firstly, as one expects, the following simple assertion holds.

Proposition 2.3. *For an arbitrary d-dimensional random vector X with $\mathbb{E}[X] = 0$ and $\mathbb{P}(X \neq 0) > 0$, we have*

$$0 < p_{d+1,X} < p_{d+2,X} < \cdots < p_{n,X} < \cdots \to 1.$$

It still holds if we only assume $p_{n,X} > 0$ for some n instead of $\mathbb{E}[X] = 0$.

The next one includes a little quantitative relation among $p_{n,X}$ and N_X.

Proposition 2.4. *For an arbitrary d-dimensional random vector X and integers $n \geq m \geq d+1$,*

$$p_{n,X} \leq \binom{n}{m} p_{m,X}, \qquad N_X \leq \frac{n}{p_{n,X}}$$

hold.

Remark 2.1. Although the estimate $N_X \leq \frac{n}{p_{n,X}}$ looks loose in general, $N_X \leq \frac{2d}{p_{2d,X}}$ is a sharp uniform bound for each dimension d up to a universal constant. Indeed, in Example 2.33 and Example 2.34 (Appendix 2.C), we prove that

$$\lim_{\varepsilon \searrow 0} \sup_{\substack{X: d\text{-dimensional} \\ p_{2d,X} < \varepsilon}} \frac{N_X p_{2d,X}}{2d} \geq \frac{1}{4}$$

holds for each positive integer d. In contrast, the other inequality $p_{n,X} \leq \binom{n}{m} p_{m,X}$ is indeed very loose and drastically improved in Proposition 2.6.

In Proposition 2.3 and 2.4, we have never used the information of dimension except for observing $p_{d+1,X} > 0$ in Proposition 2.3. However, when the distribution of X has a certain regularity, there already exists a strong result that reflects the dimensionality.

Theorem 2.5 ([172])**.** *When the distribution of X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$,*

$$p_{n,X} \leq 1 - \frac{1}{2^{n-1}} \sum_{i=0}^{d-1} \binom{n-1}{i} = \frac{1}{2^{n-1}} \sum_{i=0}^{n-d-1} \binom{n-1}{i} \tag{2.2}$$

holds for each $n \geq d+1$. The equality is attained if and only if the distribution is balanced, i.e., $\mathbb{P}(\langle c, X \rangle \leq 0) = 1/2$ holds for all the unit vectors $c \in \mathbb{R}^d$.

Wagner and Welzl [172] derived this result by showing the existence of a nonnegative continuous function h_X on $[0, 1]$ such that $h_X(t) = h_X(1-t)$, $h_X(t) \leq \frac{d+1}{2} \min\{t^d, (1-t)^d\}$ and

$$p_{n,X} = 2\binom{n}{d+1} \int_0^1 t^{n-d-1} h_X(t) \, \mathrm{d}t. \tag{2.3}$$

We shall provide an intuitive description of the function h_X. Let us consider a one-dimensional i.i.d. sequence $Y_1, Y_2, \ldots$ (also independent from $X_1, X_2, \ldots$), where each Y_i follows the uniform distribution over $(0,1)$. If we consider the $(d+1)$-dimensional random vectors $\widetilde{X}_i := (X_i, Y_i)$, then, for each n, $0 \in \operatorname{conv}\{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ is obviously equivalent to the condition that the $(d+1)$-th coordinate axis (denoted by ℓ) intersects the convex set $\widetilde{C}_n := \operatorname{conv}\{\widetilde{X}_1, \ldots, \widetilde{X}_n\} \subset \mathbb{R}^{d+1}$.

Under a certain regularity condition, there are exactly two facets (a d-dimensional face of $\widetilde{C}_n$) respectively composed of a $(d+1)$-point subset of $\{\widetilde{X}_1, \ldots, \widetilde{X}_n\}$ that intersects ℓ. Let us call them *top* and *bottom*, where the top is the facet whose intersection with ℓ has the bigger $(d+1)$-th coordinate. Let us define another random variable H as

- 0 if ℓ does not intersect $\operatorname{conv}\{\widetilde{X}_1, \ldots, \widetilde{X}_{d+1}\}$,
- otherwise the probability that 0 and $\widetilde{X}_{d+2}$ are on the same side of the hyperplane supporting $\operatorname{conv}\{\widetilde{X}_1, \ldots, \widetilde{X}_{d+1}\}$ (conditioned by $\widetilde{X}_1, \ldots, \widetilde{X}_{d+1}$).

Then, for a given realization of $\{\widetilde{X}_1, \ldots, \widetilde{X}_n\}$, the probability that $\operatorname{conv}\{\widetilde{X}_1, \ldots, \widetilde{X}_{d+1}\}$ becomes the top of $\widetilde{C}_n$ is H^{n-d-1}. As there are $\binom{n}{d+1}$ choice of (equally) possible "top," we can conclude that

$$\begin{aligned} p_{n,X} &= \mathbb{P}\Big(\ell \text{ intersects } \widetilde{C}_n\Big) \\ &= \binom{n}{d+1}\mathbb{P}\Big(\{X_1, \ldots, X_{d+1}\} \text{ is the top of } \widetilde{C}_n\Big) = \binom{n}{d+1}\mathbb{E}\big[H^{n-d-1}\big]. \end{aligned}$$

A similar observation shows $p_{n,X} = \binom{n}{d+1}\mathbb{E}\big[(1-H)^{n-d-1},\ H > 0\big]$, and so we can understand h_X as the density of a half mixture of H and $1-H$ over $\{H > 0\}$. This has been a simplified explanation of h_X. For more rigorous arguments and proofs, see Wagner and Welzl [172].

By using this "density" function, we can prove the following interesting relationship.

Proposition 2.6. *Let X be an $\mathbb{R}^d$-valued random variable with an absolutely continuous distribution. Then, for any integers $n \ge m \ge d+1$, we have*

$$\frac{1}{2^{n-m}}\frac{n(n-1)\cdots(n-d)}{m(m-1)\cdots(m-d)}p_{m,X} \le p_{n,X} \le \frac{n(n-1)\cdots(n-d)}{m(m-1)\cdots(m-d)}p_{m,X}. \qquad (2.4)$$

Proof. The right inequality is clear from (2.3). For the left inequality, by using $h_X(t) = h_X(1-t)$, we can rewrite (2.3) as

$$\begin{aligned} p_{n,X} &= \binom{n}{d+1} \int_0^1 t^{n-d-1}(h_X(t) + h_X(1-t))\,\mathrm{d}t \\ &= \binom{n}{d+1} \int_0^1 (t^{n-d-1} + (1-t)^{n-d-1})h_X(t)\,\mathrm{d}t. \end{aligned}$$

We can prove for $a \geq b \geq 0$ that $\frac{t^a+(1-t)^a}{t^b+(1-t)^b}$ attains its minimum at $t = 1/2$, e.g., by using the method of Lagrange multipliers. Accordingly, we obtain

$$\begin{aligned} \frac{p_{n,X}}{\binom{n}{d+1}} &= \int_0^1 (t^{n-d-1} + (1-t)^{n-d-1})h_X(t)\,\mathrm{d}t \\ &\geq 2^{m-n} \int_0^1 (t^{m-d-1} + (1-t)^{m-d-1})h_X(t)\,\mathrm{d}t = 2^{m-n} \frac{p_{m,X}}{\binom{m}{d+1}}, \end{aligned}$$

which is equivalent to the inequality to prove. □

Remark 2.2. The left inequality has nothing to say when n and m are large so 2^{n-m} is faster than $(n/m)^d$. However, for small n and m, it works as a nice estimate. Consider the case $n = 2d$ and $m = d+1$. Then, the proposition and the usual estimate for central binomial coefficients yield

$$p_{2d,X} \geq \frac{1}{2^{d-1}} \binom{2d}{d+1} p_{d+1,X} \geq \frac{1}{2^{d-1}} \left(\frac{d}{d+1} \frac{2^{2d}}{2\sqrt{d}} \right) p_{d+1,X} = \frac{2^d\sqrt{d}}{d+1} p_{d+1,X}.$$

This is comparable to the symmetric case, where $p_{d+1,X} = 1/2^d$ and $p_{2d,X} = 1/2$ hold.

The right inequality is an obvious improvement of the dimension-free estimate given in Proposition 2.4.

We next generalize these results to general distributions including discrete ones such as empirical measures. However, at least we have to assume $p_{d,X} = 0$. Note that it is weaker than the condition that X has an absolutely continuous distribution, as it is satisfied with usual empirical measures (see Proposition 2.8).

From smoothing arguments, we obtain the following generalization of inequalities (2.2) and (2.4).

Theorem 2.7. *Let X be an arbitrary d-dimensional random vector with $p_{d,X} = 0$. Then, for any $n \geq m \geq d+1$, the following inequalities hold:*

$$p_{n,X} \leq 1 - \frac{1}{2^{n-1}} \sum_{i=0}^{d-1} \binom{n-1}{i}, \qquad \frac{1}{2^{n-m}} \frac{\binom{n}{d+1}}{\binom{m}{d+1}} p_{m,X} \leq p_{n,X} \leq \frac{\binom{n}{d+1}}{\binom{m}{d+1}} p_{m,X}.$$

We should remark that $p_{d,X} = 0$ is naturally satisfied with (centered) empirical measures.

Proposition 2.8. *Let μ be an absolutely continuous probability distribution on $\mathbb{R}^d$ and $Y_1, Y_2, \ldots$ be an i.i.d. samplings from μ. Then, with probability one, for each $M \geq d+1$, distributions*

$$\mu_M := \frac{1}{M} \sum_{i=1}^{M} \delta_{Y_i} \quad \textit{and} \quad \widetilde{\mu}_M := \frac{1}{M} \sum_{i=1}^{M} \delta_{Y_i - \frac{1}{M} \sum_{j=1}^{M} Y_j}$$

satisfy $p_{d,\mu_M} = p_{d,\widetilde{\mu}_M} = 0$. We also have $p_{d,\mu_M} = 0$ for $1 \leq M \leq d$; it only requires $p_{d,\mu} = 0$ rather than absolute continuity.

2.3 Uniform bounds of $p^{\varepsilon}_{n,X}$ via the relaxed Tukey depth

We have not used any quantitative assumption on the distribution of X in the previous section. In this section, however, we shall evaluate $p_{n,X}$ and its ε-approximation version by using the Tukey depth and its relaxation. We shall fix an arbitrarily real inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^d$, and use the induced norm $\| \cdot \|$ and the notation $\operatorname{dist}(x, A) := \inf_{a \in A} \|x - a\|$ for an $x \in \mathbb{R}^d$ and $A \subset \mathbb{R}^d$.

For a d-dimensional random vector X and $\theta \in \mathbb{R}^d$, define an ε-relaxation version of the Tukey depth by

$$\alpha^{\varepsilon}_X(\theta) := \inf_{\|c\|=1} \mathbb{P}(\langle c, X - \theta \rangle \leq \varepsilon).$$

We also define, for a positive integer n,

$$p^{\varepsilon}_{n,X}(\theta) := \mathbb{P}(\operatorname{dist}(\theta, \operatorname{conv}\{X_1, \ldots, X_n\}) \leq \varepsilon),$$

where $X_1, \ldots, X_n$ are independent copies of X. Note that $p_{n,X} = p^0_{n,X}$. Although we regard them as functions of θ in Section 2.5, we only treat the case $\theta = 0$ and omit the argument θ in this section.

Proposition 2.9. *Let X be a d-dimensional random vector with an absolutely continuous distribution with respect to the Lebesgue measure. Then, for each $\varepsilon \geq 0$ and positive integer $n \geq d+1$, we have*

$$1 - p_{n,X}^{\varepsilon} \leq \frac{n(1-\alpha_X^{\varepsilon})}{n-d}(1 - p_{n-1,X}^{\varepsilon}).$$

Before going into details of quantitative results, we note the following equivalence of the positivity of α_X^{ε} and $p_{n,X}^{\varepsilon}$ as an immediate consequence of this assertion.

Proposition 2.10. *Let X be an arbitrary d-dimensional random vector and let $\varepsilon \geq 0$. Then, $p_{n,X}^{\varepsilon} > 0$ for some $n \geq 1$ implies $\alpha_X^{\varepsilon} > 0$. Reciprocally, $\alpha_X^{\varepsilon} > 0$ implies $p_{n,X}^{\varepsilon} > 0$ for all $n \geq d+1$.*

Let us prove Proposition 2.9. We give its combinatorial proof here since it is the primary technical contribution of this chapter.

Proof of Proposition 2.9. Let $m \geq d$ be an integer. We first consider the quantity $q_m := 1 - p_{m,X}^{\varepsilon}$. Let A_m be the event given by $\operatorname{dist}(0, \operatorname{conv}\{X_i\}_{i=1}^m) > \varepsilon$. Also, let B_m be the event that $\{X_1, \ldots, X_m\}$ is in general position. Then, we have $\mathbb{P}(B_m) = 1$ and $q_m = \mathbb{P}(A_m \cap B_m)$.

Under the event $A_m \cap B_m$, we have a unique point $h_m \in \operatorname{conv}\{X_i\}_{i=1}^m$ that minimizes $\|h_m\|$. Let H_m be the open halfspace defined by $H_m := \{x \in \mathbb{R}^d \mid \langle x - h_m, h_m \rangle > 0\}$. Then, the boundary ∂H_m is the hyperplane going through h_m and perpendicular to h_m. From the general-position assumption, there are at most d points in $\{X_i\}_{i=1}^m \cap \partial H_m$. Let I_m be the set of indices i satisfying $X_i \in \partial H_m$, then I_m is a random subset of $\{1, \ldots, m\}$ with $1 \leq |I_m| \leq d$ under the event $A_m \cap B_m$. Note also that $X_i \in H_m$ for each $i \in \{1, \ldots, m\} \setminus I_m$. For simplicity, define $I_m = \emptyset$ for the event $(A_m \cap B_m)^c$.

As I_m is a random set determined uniquely, we can decompose the probability $\mathbb{P}(A_m \cap B_m)$ as follows by symmetry:

$$q_m = \mathbb{P}(A_m \cap B_m) = \sum_{k=1}^{d} \binom{m}{k} \mathbb{P}(I_m = \{1, \ldots, k\}).$$

Hence, we want to evaluate the probability $\mathbb{P}(I_m = \{1, \ldots, k\})$. Note that we can similarly define h_k as the unique point in $\operatorname{conv}\{X_i\}_{i=1}^k$ that minimizes the distance from the origin. Then, H_k is the open halfspace $H_k = \{x \in \mathbb{R}^d \mid \langle x - h_k, h_k\rangle > 0\}$. Then, we have

$$\begin{aligned}\mathbb{P}(I_m = \{1, \dots, k\}) &= \mathbb{E}\left[\mathbb{1}_{\{\|h_k\|>\varepsilon,\ \mathrm{conv}\{X_i\}_{i=1}^k \subset \partial H_k\}} \prod_{j=k+1}^{m} \mathbb{P}\bigl(X_j \in H_k \mid \{X_i\}_{i=1}^k\bigr)\right] \\ &= \mathbb{E}\Bigl[\mathbb{1}_{\{\|h_k\|>\varepsilon,\ \mathrm{conv}\{X_i\}_{i=1}^k \subset \partial H_k\}} \mathbb{P}\bigl(X' \in H_k \mid \{X_i\}_{i=1}^k\bigr)^{m-k}\Bigr],\end{aligned}$$

where X' is a copy of X independent from $X_1, X_2, \dots$. As $\mathbb{P}\bigl(X' \in H_k \mid \{X_i\}_{i=1}^k\bigr) \le 1 - \alpha_X^\varepsilon$ under the event $\{\|h_k\| > \varepsilon,\ \mathrm{conv}\{X_i\}_{i=1}^k \subset \partial H_k\}$, we have

$$\begin{aligned}\mathbb{P}(I_{m+1} = \{1, \dots, k\}) &= \mathbb{E}\Bigl[\mathbb{1}_{\{\|h_k\|>\varepsilon,\ \mathrm{conv}\{X_i\}_{i=1}^k \subset \partial H_k\}} \mathbb{P}\bigl(X' \in H_k \mid \{X_i\}_{i=1}^k\bigr)^{m+1-k}\Bigr] \\ &\le (1 - \alpha_X^\varepsilon)\mathbb{P}(I_m = \{1, \dots, k\}).\end{aligned}$$

Therefore, we have

$$\begin{aligned}q_{m+1} &= \sum_{k=1}^{d} \binom{m+1}{k} \mathbb{P}(I_{m+1} = \{1, \dots, k\}) \\ &= \sum_{k=1}^{d} \frac{m+1}{m+1-k}\binom{m}{k}(1 - \alpha_X^\varepsilon)\mathbb{P}(I_m = \{1, \dots, k\}) \\ &\le \frac{(m+1)(1-\alpha_X^\varepsilon)}{m+1-d} q_m.\end{aligned}$$

By letting $n = m + 1$, we obtain the conclusion. □

If we define $g_{d,n}(\alpha)$ for $\alpha \in [0, 1]$ by $g_{d,n} := 1$ for $n = 1, \dots, d$ and

$$g_{d,n}(\alpha) := \min\left\{1, \frac{n(1-\alpha)}{n-d} g_{d,n-1}(\alpha)\right\} \tag{2.5}$$

for $n = d + 1, d + 2, \dots$, we clearly have $1 - p_{n,X}^\varepsilon \le g_{d,n}(\alpha_X^\varepsilon)$ from Proposition 2.9 for a d-dimensional X with density. We can actually generalize this to any X.

Lemma 2.11. *Let X be an arbitrary d-dimensional random vector. Then, for each $\varepsilon \ge 0$ and positive integer n, we have $1 - p_{n,X}^\varepsilon \le g_{d,n}(\alpha_X^\varepsilon)$.*

For a special choice $n = \lceil 3d/\alpha \rceil$, we obtain the following main result from a concrete estimate of $g_{d,n}(\alpha)$.

Theorem 2.12. *Let X be an arbitrary d-dimensional random vector. Then, for each $\varepsilon \geq 0$ and positive integer $n \geq 3d/\alpha_X^\varepsilon$, we have*

$$p_{n,X}^\varepsilon > 1 - \frac{1}{2^d}.$$

We also know the following assertion as a consequence of Theorem 2.12.

Theorem 2.13. *Let X be an arbitrary d-dimensional random vector. Then, we have*

$$\frac{1}{2\alpha_X} \leq N_X \leq \left\lceil \frac{3d}{\alpha_X} \right\rceil.$$

Proof. The right inequality is an immediate consequence of Theorem 2.12. To prove the left one, let n be a positive integer satisfying $\frac{1}{2n} > \alpha_X$. Then, there exists a vector $c \in \mathbb{R}^d \setminus \{0\}$ such that $\mathbb{P}(\langle c, X \rangle \leq 0) < \frac{1}{2n}$. Then, for $X_1, X_2, \ldots, X_n$ (i.i.d. copies of X), we have

$$p_{n,X} = \mathbb{P}(0 \in \operatorname{conv}\{X_1, \ldots, X_n\}) \leq \mathbb{P}\left(\bigcup_{i=1}^n \{\langle c, X_i \rangle \leq 0\}\right) \leq n\mathbb{P}(\langle c, X \rangle \leq 0) < \frac{1}{2}.$$

Therefore, N_X must satisfy $\frac{1}{2N_X} \leq \alpha_X$. □

Remark 2.3. The above theorem states that $1/2 \leq \alpha_X N_X \leq 3d + 1$. This evaluation for $\alpha_X N_X$ is indeed tight up to a universal constant. For example, if X is a d-dimensional standard Gaussian, we have $\alpha_X = \frac{1}{2}$ and $N_X = 2d$, so $\alpha_X N_X = d$. Moreover, for a small $\varepsilon \in (0, 1)$, if we consider $X = (X^1, \ldots, X^d)$ such that

- $\mathbb{P}\big(X^d = 1\big) = \varepsilon$ and $\mathbb{P}\big(X^d = -1\big) = 1 - \varepsilon$,
- $(X^1, \ldots, X^{d-1})|_{X^d=1}$ is a standard Gaussian,
- $X^1 = \cdots = X^{d-1} = 0$ if $X^d = -1$,

then we can see $\alpha_X = \varepsilon/2$ and $N_X = \Omega((d-1)/\varepsilon)$ as $(0, \ldots, 0, 1)$ has to be in the convex hull of samples to include the origin in it. Hence the bound $\alpha_X N_X = \mathcal{O}(d)$ is sharp even for a small α_X.

On the contrary,

$$\inf_{X:d\text{-dimensional}} \alpha_X N_X \leq 2$$

holds (even when requiring $p_{d,X} = 0$) for each positive integer d from Example 2.33 and Example 2.34 in the appendix (Section 2.C).

We complete this section with a stronger version of Proposition 2.9 only for $\varepsilon = 0$. Indeed, by summing up the following inequality, we can immediately obtain the $\varepsilon = 0$ case in Proposition 2.9.

Proposition 2.14. *Let X be a d-dimensional random vector with an absolutely continuous distribution with respect to the Lebesgue measure. Then,*

$$p_{n+1,X} - p_{n,X} \le \frac{n(1-\alpha_X)}{n-d}(p_{n,X} - p_{n-1,X})$$

holds for all $n \ge d+1$.

2.4 Bounds of N_X via Berry–Esseen theorem

In this section, we discuss moment-based bounds of N_X for a centered X, which are of particular interest from the randomized measure reduction (see Section 1.1).

Although Theorem 2.13 has strong generality, we have little information about the Tukey depth α_X in many situations. Indeed, approximately computing the Tukey depth itself is an important and difficult problem [36, 184]. However, if we limit the argument to a centered X, we can obtain various moment-based bounds as shown below. In this section, we use the usual Euclidean norm $\|\cdot\|_2$ given by $\|x\|_2 = \sqrt{x^\top x}$ for simplicity.

Let X be a d-dimensional centered random vector whose covariance matrix $V := \mathbb{E}\left[XX^\top\right]$ is nonsingular. We also define $V^{-1/2}$ as the positive-definite square root of V^{-1}. Then, for each unit vector $c \in \mathbb{R}^d$ (namely $\|c\|_2 = 1$), we have

$$\mathbb{E}\left[(c^\top V^{-1/2}X)^2\right] = \mathbb{E}\left[c^\top V^{-1/2}XX^\top V^{-1/2}c\right] = \mathbb{E}\left[c^\top c\right] = 1, \tag{2.6}$$

We have the following simple result for a bounded X.

Proposition 2.15. *Let X be a centered d-dimensional random vector with nonsingular covariance matrix V. If $\|V^{-1/2}X\|_2 \le B$ holds almost surely for a positive constant B, then we have*

$$\alpha_X \ge \frac{1}{2B^2}, \qquad N_X \le \left\lceil 6dB^2 \right\rceil.$$

Let us consider the unbounded case. The Berry–Esseen theorem evaluates the speed of convergence in the central limit theorem [19, 48]. The following is a recent result with an explicit small constant.

Theorem 2.16 ([95])**.** *Let Y be a random variable with $\mathbb{E}[Y] = 0$, $\mathbb{E}[Y^2] = 1$, and $\mathbb{E}[|Y|^3] < \infty$, and let $Y_1, Y_2, \ldots$ be independent copies of Y. Also, let Z be a one-dimensional standard Gaussian. Then, we have*

$$\left|\mathbb{P}\!\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}} \le x\right) - \mathbb{P}(Z \le x)\right| \le \frac{0.4784\,\mathbb{E}[|Y|^3]}{\sqrt{n}}$$

for arbitrary $x \in \mathbb{R}$ and $n \ge 1$.

We can apply the Berry–Esseen theorem for evaluating the probability $\mathbb{P}\big(c^\top S_n \le 0\big)$ from (2.6), where S_n is the normalized i.i.d. sum $\frac{1}{\sqrt{n}} V^{-1/2}(X_1 + \cdots + X_n)$. By elaborating this idea, we obtain the following bound of N_X.

Theorem 2.17. *Let X be a centered d-dimensional random vector with nonsingular covariance matrix V. Then, we have*

$$N_X \le 17d \left(1 + \frac{9}{4} \sup_{c \in \mathbb{R}^d, \|c\|_2 = 1} \mathbb{E}\left[\left|c^\top V^{-1/2} X\right|^3\right]^2\right).$$

Remark 2.4. The bound in Theorem 2.17 is sharp up to a constant as a uniform bound in terms of $\mathbb{E}\left[\left|c^\top V^{-1/2} X\right|^3\right]$. Indeed, if X is d-dimensional standard Gaussian, then $\mathbb{E}\left[\left|c^\top V^{-1/2} X\right|^3\right] = \frac{2\sqrt{2}}{\sqrt{\pi}}$ holds for all $\|c\|_2 = 1$ while $N_X = 2d$, so

$$\sup_{c \in \mathbb{R}^d, \|c\|_2 = 1} \mathbb{E}\left[\left|c^\top V^{-1/2} X\right|^3\right]^{-2} N_X = \frac{\pi}{4} d.$$

From Theorem 2.17, we also obtain several looser but more tractable bounds.

Corollary 2.18. *Let X be a centered d-dimensional random vector with nonsingular covariance matrix V. N_X can be bounded as*

$$N_X \le 17d \left(1 + \frac{9}{4} \min\left\{\mathbb{E}\left[\left\|V^{-1/2} X\right\|_2^3\right]^2,\ \mathbb{E}\left[\left\|V^{-1/2} X\right\|_2^4\right]\right\}\right).$$

Remark 2.5. In the order notation, the first bound in this corollary states

$$N_X = \mathcal{O}\left(d\,\mathbb{E}\left[\left\| V^{-1/2}X \right\|_2^3 \right]^2 \right).$$

This estimate is also sharp up to $\mathcal{O}(d)$ factor in the sense that we can prove

$$\sup \left\{ \frac{N_X}{\mathbb{E}\left[\left\| V^{-1/2}X \right\|_2^3 \right]^2} \;\middle|\; \begin{array}{c} X \text{ is } d\text{-dimensional, } \mathbb{E}[X] = 0, \\ V = \mathbb{E}\left[XX^\top\right] \text{ is nonsingular, } \mathbb{E}\left[\left\| V^{-1/2}X \right\|_2^3 \right] < \infty \end{array} \right\} \geq \frac{1}{2}$$

for each positive integer d. For proof of this fact, see Example 2.33 and Example 2.34 in the appendix (Section 2.C).

We finally remark that there are multivariate versions of the Berry–Esseen theorem [182, 140] and we can use them to derive a bound of N_X in a different approach which does not use α_X. However, their bounds only give the estimate

$$N_X = \mathcal{O}\left(d^{7/2}\mathbb{E}\left[\left\| V^{-1/2}X \right\|_2^3 \right]^2 \right), \tag{2.7}$$

which is far worse than the bounds obtained in Theorem 2.17 and Corollary 2.18. However, it is notable that this approach from multidimensional Berry–Esseen formulas is applicable to *non-identical* X_i's if the second and third moments are uniformly bounded, while the combinatorial approach based on α_X seems to be fully exploiting the i.i.d. assumption. Therefore, we provide the details of this alternative approach in the appendix (Section 2.B).

2.5 Deterministic interior body of random polytopes

For each $\alpha > 0$, define a deterministic set defined by the level sets of Tukey depth

$$K^\alpha(X) := \{\theta \in \mathbb{R}^d \mid \alpha_X(\theta) \geq \alpha\}.$$

This set is known to be compact and convex [144]. We can also naturally generalize this set for the ε-relaxation of Tukey depth, and the generalization also satisfies the following:

Proposition 2.19. *Let X be a d-dimensional random vector. Then, for each $\varepsilon \geq 0$ and $\alpha > 0$, the set $\{\theta \in \mathbb{R}^d \mid \alpha_X^\varepsilon(\theta) \geq \alpha\}$ is compact and convex, and satisfies*

$$\{\theta \in \mathbb{R}^d \mid \alpha_X^\varepsilon(\theta) \geq \alpha\} \supset \{\theta \in \mathbb{R}^d \mid \mathrm{dist}(\theta, K^\alpha(X)) \leq \varepsilon\}.$$

Remark 2.6. Note that the inclusion stated in Proposition 2.19 can be strict. For example, if X is a d-dimensional standard Gaussian, $K^\alpha(X)$ is empty for each $\alpha > 1/2$, but the ε-relaxation of Tukey depth can be greater than $1/2$ for $\varepsilon > 0$.

From this proposition, we can naturally generalize the arguments given in this section to the ε-relaxation case; natural interior bodies of ε-neighborhood of $\mathrm{conv}\{X_1, \ldots, X_n\}$ are given by the ε-relaxation of Tukey depth. However, to keep the notation simple, we only treat $K^\alpha(X)$ the interior body of the usual convex hull in the following.

We next prove that the polar body $\big(\widetilde{K}^\alpha(X)\big)^\circ$ used in Guédon et al. [60], which we have introduced in Section 2.1.2, is essentially the same as $K^\alpha(X)$ in their setting, i.e., when X is symmetric. Recall that $\widetilde{K}^\alpha(X)$ is defined as

$$\widetilde{K}^\alpha(X) = \{s \in \mathbb{R}^d \mid \mathbb{P}(\langle s, X\rangle \geq 1) \leq \alpha\}.$$

Note that the following proposition is not surprising if we go back to the original background of $\widetilde{K}^\alpha$ [150], where X is uniform from some deterministic convex set, and recent research on its relation to the Tukey depth [124].

Proposition 2.20. *Let X be a d-dimensional symmetric random vector. Then, for each $\alpha \in (0, 1/2)$, we have*

$$\{\theta \in \mathbb{R}^d \mid \alpha_X(\theta) > \alpha\} \subset \big(\widetilde{K}^\alpha(X)\big)^\circ \subset K^\alpha(X).$$

We are going to prove the extension of Theorem 2.2 by finding a finite set of points whose convex hull approximates $K^\alpha(X)$. The following statement is essentially well-known [138, 10], but we give the precise statement and a brief proof (Appendix 2.A.14) for completeness.

Proposition 2.21. *Let K be a compact and convex subset of $\mathbb{R}^d$ such that $K = -K$. Then, for each $\varepsilon \in (0, 1)$, there is a finite set $A \subset \mathbb{R}^d$ such that*

$$(1-\varepsilon)K \subset \mathrm{conv}\, A \subset K, \qquad |A| \leq \left(1 + \frac{2}{\varepsilon}\right)^d.$$

Theorem 2.22. *Let X be an arbitrary symmetric d-dimensional random vector, and let $\alpha, \delta, \varepsilon \in (0, 1)$. If a positive integer n satisfies*

$$n \geq \frac{2d}{\alpha} \max \left\{ \frac{\log(1/\delta)}{d} + \log \frac{1}{\varepsilon},\ 6 \right\},$$

then we have, with probability at least $1 - \delta$,

$$\operatorname{conv}\{X_1, \dots, X_n\} \supset (1 - \varepsilon) K^\alpha(X),$$

where $X_1, X_2, \dots$ are independent copies of X.

Remark 2.7. Although the bound given in Theorem 2.22 requires $n \geq 12d/\alpha$, it can be loosened for moderate δ and ε. For example, if we want to obtain a bound for the case $\delta = \varepsilon = 1/2$, then we can prove $n \geq 5d/\alpha$ to be sufficient by using the bound in Proposition 2.24. Moreover, we should note that we have used the assumption that X is symmetric only to prove that $K^\alpha(X)$ is symmetric (so that we can use Proposition 2.21). If we take a symmetric convex subset $K \subset K^\alpha(X)$, we can prove a similar inclusion statement for K even for a nonsymmetric X.

If we want a generalized version of Theorem 2.2, we can prove the following:

Corollary 2.23. *Let X be an arbitrary d-dimensional symmetric random vector. Let $\beta \in (0, 1)$ and set $\alpha = (en/d)^{-\beta}$. Then, there exists an absolute constant $c > 0.45$ such that, for each integer n satisfying $n \geq (12e^\beta)^{1/(1-\beta)} d$, we have*

$$\operatorname{conv}\{X_1, \dots, X_n\} \supset \frac{1}{2} K^\alpha(X)$$

with probability at least $1 - \exp(-ce^{-\beta} n^{1-\beta} d^\beta)$, where $X_1, X_2, \dots$ are independent copies of X.

2.6 Application

We discuss the implications of the results of this chapter in two parts. The first part discusses the use of the bounds we gave on $p_{n,X}$, while the second part gives the implication of N_X's bounds on the randomized cubature construction.

2.6.1 Bounds of $p_{n,X}$

Firstly, the inequality between $p_{n,X}$ and $p_{m,X}$ given in Proposition 2.6 provides the inequality

$$p_{2d,X} \geq \frac{2^d\sqrt{d}}{d+1} p_{d+1,X} \tag{2.8}$$

as it is mentioned in Remark 2.2.

Measure reduction. Consider a discrete (probability) measure $\mu = \sum_{x\in\mathcal{X}} w_x\delta_x$ for a finite subset of $\mathcal{X} \subset \mathbb{R}^d$. In Cosentino et al. [34], randomized algorithms for constructing a convex combination satisfying $\mathbb{E}_{X\sim\mu}[X] = \sum_{i=1}^{d+1} \lambda_i x_i$ $(x_i \in \mathcal{X})$, whose existence is assured by Tchakaloff's theorem [162, 12], are considered. As a basic algorithm, the authors consider the following scheme:

(a.1) Randomly choose d points $A = \{x_1, \dots, x_d\}$ from $\mathcal{X}$.

(a.2) For each $x \in \mathcal{X} \setminus A$, determine if $\mathbb{E}_{X\sim\mu}[X] \in \operatorname{conv}(A \cup \{x\})$ or not, and finish the algorithm and return $A \cup \{x\}$ if it holds.

(a.3) Go back to (a.1).

Although we can execute the decision for each x in (a.2) with $\mathcal{O}(d^2)$ computational cost with an $\mathcal{O}(d^3)$ preprocessing for a fixed A, the overall expected computational cost until the end of the algorithm is at least $\Omega(d^2/p_{d+1,X})$ under some natural assumption on μ (see Proposition 2.8).

However, we can also consider the following naive procedure:

(b.1) Randomly choose $2d$ points $B = \{x_1, \dots, x_{2d}\}$ from $\mathcal{X}$.

(b.2) Return B if $\mathbb{E}_{X\sim\mu}[X] \in \operatorname{conv} B$, and go back to (b.1) if not.

By using an LP solver with the simplex method we can execute (b.2) in (empirically) $\mathcal{O}(d^3)$ time [133, 151]. Hence the overall computational cost can be heuristically bounded above by $\mathcal{O}(d^3/p_{2d,X})$, which is faster than the former by $\Omega\big(d^{-3/2}2^d\big)$ from the evaluation in (2.8). Note also that we have rigorously polynomial bounds via other LP methods (e.g., an infeasible-interior-point method; [120]), and so the latter scheme is preferable even in worst-case when the dimension d becomes large.

Relation between two depths. We can also deduce an inequality between two depth concepts in statistics. As mentioned in Introduction, for a random vector $X \in \mathbb{R}^d$, $p_{d+1,X}$ is called the simplicial depth whereas α_X is the Tukey depth of the origin with respect to X.

Naively, we have $\alpha_X \geq \frac{p_{n,X}}{n}$ for each n, so $\alpha_X \geq \frac{p_{d+1,X}}{d+1}$ holds. However, by using (2.8) here, we obtain a sharper estimate

$$\alpha_X \geq \frac{p_{2d,X}}{2d} \geq \frac{1}{2d}\frac{2^d\sqrt{d}}{d+1}p_{d+1,X} \geq \frac{2^{d-1}}{\sqrt{d}(d+1)}p_{d+1,X}.$$

In contrast, deriving a nontrivial upper bound of α_X in terms of $p_{d+1,X}$ still seems difficult.

2.6.2 Bounds of N_X

Secondly, we give applications of the bounds of N_X given in Section 2.4.

Random trigonometic cubature. Consider a d-dimensional random vector

$$X = (\cos\theta, \dots, \cos d\theta)^\top \in \mathbb{R}^d$$

for a positive integer d, where θ is a uniform random variable over $(-\pi, \pi)$. Then, from an easy computation, we have $V := \mathbb{E}\bigl[XX^\top\bigr] = \frac{1}{2}I_d$, and so we obtain $\|V^{-1/2}X\|^2 \leq 2d$ almost surely. Therefore, from Proposition 2.15, we have $N_X \leq 1 + 12d^2$. This example is equivalent to a random construction of the so-called Gauss–Chebyshev quadrature [114, Chapter 8]. Although we can have a bound for the number of observations required in a random construction as above, concrete constructions with fewer points are already known.

Deriving a bound for random construction of cubature without any known deterministic construction, such as cubature on Wiener space [110, 68] with general degree, which is more important, is addressed in the next chapter by using the concept of hypercontractivity.

Beyond naive cubature construction. Recall the cubature construction problem described in Section 1.1. We consider a random variable of the form $X = \boldsymbol{f}(Y)$, where Y is a random variable on some topological space $\mathcal{X}$ and $\boldsymbol{f} = (f_1, \dots, f_d)^\top : \mathcal{X} \to \mathbb{R}^d$ is a d-dimensional vector valued integrable function. Our aim is to find points $y_1, \dots, y_{d+1} \in \mathcal{X}$ and weights $w_1, \dots, w_{d+1} \geq 0$ whose total is one such that

$$\mathbb{E}[\boldsymbol{f}(Y)] = \sum_{j=1}^{d+1} w_j \boldsymbol{f}(y_j). \tag{2.9}$$

A naive algorithm proposed by Hayakawa [66] was to generate independent copies $Y_1, Y_2, \dots$ of Y and choose y_j from these random samples. Without any knowledge of N_X, the algorithm would be of the form

(c.1) Take $k = 2d$.

(c.2) Randomly generate Y_i up to $i = k$ and determine if (2.9) can be satisfied with $y_j \in \{Y_i\}_{i=1}^k$ by using an LP solver.

(c.3) If we find a solution, stop the algorithm. Otherwise, go to (c.2) after replacing k with $2k$.

This procedure ends at $k \leq 2N_X(\mathbb{E}[X])$ with probability over half. We can then heuristically estimate the computational cost by $\Theta(C(d, N_X(\mathbb{E}[X])))$, where we denote by $C(d, n)$ the computational complexity of a linear programming problem finding the solution of (2.9) from n sample points. Empirically, this is estimated as $\Omega(d^2 n)$ or more when we use the simplex method [151].

However, our analysis on N_X via the Berry–Esseen bound tells us the possibility of an alternative (Algorithm 2.1).

Algorithm 2.1 Randomized cubature construction for recombination

```
Input: An integer $\ell \geq 2$
Output: $(w_1, y_1), \dots, (w_n, y_n) \in \mathbb{R}_{\geq 0} \times \mathcal{X}$ satisfying $\sum_{j=1}^n w_j = 1$ and $\mathbb{E}[X] = \sum_{j=1}^n w_j \boldsymbol{f}(y_j)$
Initialize:
    $x_1, \dots, x_{\ell d}, z_1, \dots, z_{\ell d}$ : vectors in $\mathbb{R}^d$, $k \leftarrow 0$
for $i = 1, \dots, \ell d$ do
    Sample $Y_i$
    $x_i \leftarrow \boldsymbol{f}(Y_i)$
end for
while $\mathbb{E}[X] \notin \mathrm{conv}\{x_1, \dots, x_{\ell d}\}$ do
    for $i = 1, \dots, \ell d$ do
        $z_i \leftarrow 0$ (as an $\mathbb{R}^d$ vector)
    end for
    for $j = 2^k, \dots, 2^{k+1} - 1$ do
        for $i = 1, \dots, \ell d$ do
            Sample $Y_{j\ell d+i}$
            $z_i \leftarrow z_i + 2^{-k} \boldsymbol{f}(Y_{j\ell d+i})$
        end for
    end for
    for $i = 1, \dots, \ell d$ do
        $x_i \leftarrow (x_i + z_i)/2$
    end for
    $k \leftarrow k + 1$
end while
Take $x_{i_1}, \dots, x_{i_{d+1}}$ and $\lambda_1, \dots, \lambda_{d+1}$ such that $\mathbb{E}[X] = \sum_{m=1}^{d+1} \lambda_m x_{i_m}$ by solving an LP
Return $(2^{-k}\lambda_m, Y_{j\ell d+i_m})$ for $(j, m) \in \{0, \dots, 2^k - 1\} \times \{1, \dots, d+1\}$
```

Although the pseudocode may seem a little long, this just uses ℓd random vectors of the form $n^{-1}(X_1 + \cdots + X_n)$ as the possible vertices of the convex combination, which is used for deriving bounds of N_X in Section 2.4. After executing Algorithm 2.1, we can use any algorithm for deterministic measures (typically called recombination; [103, 163, 111]) to obtain an actual $d + 1$ points cubature rule, whose time complexity is rigorously bounded by $\mathcal{O}\bigl(kd^3 + 2^k d^2\bigr)$ by using the final value of k in the above algorithm.

As we can carry out Algorithm 2.1 within $\mathcal{O}\bigl(2^k \ell d^2 + kC(d, \ell d)\bigr)$, the overall computational cost is $\mathcal{O}\bigl(kC(d, \ell d) + 2^k \ell d^2\bigr)$. Then we heuristically have the bound

$\mathcal{O}\big(k\ell d^3 + 2^k\ell d^2\big)$ for a small ℓ. By using the number $N = 2^k\ell d$, which is the number of randomly generated copies of Y, this cost is rewritten as $\mathcal{O}(\log(N/\ell d)\ell d^3 + Nd)$.

As our bound for $N_X(\mathbb{E}[X])$ in Theorem 2.17 is applicable for this N because of the use of Berry–Esseen type estimate ($\ell = 17$ is used in the proof), we can also give an estimate for this alternative algorithm. If the N is not as large as $\Omega(dN_X(\mathbb{E}[X]))$ for an appropriate choice of ℓ, we indeed have a better scheme, though the comparison itself may be a nontrivial problem in general. In any event, the fact that we can avoid solving a large LP problem is an obvious advantage.

2.7 Concluding remarks

In this chapter, we have investigated inequalities regarding $p_{n,X}$, N_X and α_X, which is motivated by the fields of numerical analysis, data science, statistics and random matrix. We generalized the existing inequalities for $p_{n,X}$ in Section 2.2. After pointing out that the convergence rate of $p_{n,X}$ is determined by α_X in Section 2.3 with introduction of ε-relaxation of both quantities, we proved that N_X and $1/\alpha_X$ are of the same magnitude up to an $\mathcal{O}(d)$ factor in Theorem 2.13. We also gave estimates of N_X based on the moments of X in Section 2.4 by using Berry–Esseen type bounds. Although arguments have been based on whether a given vector is included in the random convex polytope $\operatorname{conv}\{X_1, \ldots, X_n\}$, in Section 2.4, we extended our results to the analysis of deterministic convex bodies included in the random convex hull, which immediately led to a technical improvement on a result of the random matrix community. We finally discussed several implications of our results on application in Section 2.6.

Appendix for Chapter 2

2.A Proofs

2.A.1 Proof of Proposition 2.3

Proof. For the proof of $p_{2d,X} > 0$, see, e.g., Hayakawa [66]. From this and Carathéodory's theorem, we also have $p_{d+1,X} > 0$. We clearly have $p_{n+1,X} \geq p_{n,X}$ for each $n \geq d + 1$.

The strict inequality also seems trivial, but we prove this for completeness. Assume $p_{n+1,X} = p_{n,X}$ for some n. This implies that $0 \not\in \mathrm{conv}\{X_i\}_{i=1}^n \Rightarrow 0 \not\in \mathrm{conv}\{X_i\}_{i=1}^{n+1}$ holds almost surely. By symmetry, for any $J \subset \{1,\dots,n+2\}$ with $|J| = n+1$, $0 \not\in \mathrm{conv}\{X_i\}_{i=1}^{n+1} \Rightarrow 0 \not\in \mathrm{conv}\{X_i\}_{i\in J}$ holds almost surely. Therefore, we have $0 \not\in \mathrm{conv}\{X_i\}_{i=1}^n \Rightarrow 0 \not\in \mathrm{conv}\{X_i\}_{i=1}^{n+2}$ with probability one. By repeating this argument, we obtain

$$0 \not\in \mathrm{conv}\{X_1,\dots,X_n\} \Longrightarrow 0 \not\in \mathrm{conv}\{X_1,\dots,X_{n+d+1}\} \Longrightarrow 0 \not\in \mathrm{conv}\{X_{n+1},\dots,X_{n+d+1}\}$$

with probability one, but this is only possible when $\mathbb{P}(0 \not\in \mathrm{conv}\{X_1,\dots,X_n\}) = 0$ as $p_{d+1,X} > 0$ and the variables $X_{n+1},\dots,X_{n+d+1}$ are independent from the others. This is of course impossible from the assumption $\mathbb{P}(X \neq 0) > 0$ (there exists a unit vector $c \in \mathbb{R}^d$ such that $\mathbb{P}(\langle c, X\rangle > 0) > 0$), so we finally obtain $p_{n,X} < p_{n+1,X}$.

Proving $p_{n,X} \to 1$ is also easy. From the independence, we have

$$\begin{aligned} p_{m(d+1),X} &= 1 - \mathbb{P}\big(0 \not\in \mathrm{conv}\{X_1,\dots,X_{m(d+1)}\}\big) \\ &\ge 1 - \mathbb{P}\left(\bigcap_{k=1}^m \{0 \not\in \mathrm{conv}\{X_{(k-1)(d+1)+1},\dots,X_{k(d+1)}\}\}\right) \\ &= 1 - (1 - p_{d+1,X})^m \to 1 \qquad (m \to \infty). \end{aligned}$$

This leads to the conclusion combined with the monotonicity of $p_{n,X}$.

Note that we have used the condition $\mathbb{E}[X] = 0$ only to ensure $p_{d+1} > 0$. Hence the latter statement readily holds from the same argument. □

2.A.2 Proof of Proposition 2.4

Proof. Let M be the number of m-point subsets of $\{X_1,\dots,X_n\}$ whose convex hull contains 0. Then, we have

$$\mathbb{E}[M] = \sum_{\substack{J\subset\{1,\dots,n\} \\ |J|=m}} \mathbb{P}(0 \in \mathrm{conv}\{X_i\}_{i\in J}) = \binom{n}{m} p_{m,X}.$$

As $p_{n,X} = \mathbb{P}(M \ge 1) \le \mathbb{E}[M]$, we obtain the first inequality.

For the second part, we carry out the following rough estimate: For the minimum integer k satisfying $(1-p_{n,X})^k \le 1/2$, we have $N_X \le kn$. If $p_{n,X} \ge 1/2$

holds, then $N_X \leq n$ immediately holds. Thus it suffices to prove $k \leq \left\lceil \frac{1-p_{n,X}}{p_{n,X}} \right\rceil$ when $p_{n,X} < 1/2$. Indeed, by the monotonicity of $(1+1/x)^x$ over $x > 0$, we have

$$\left(\frac{1}{1-p_{n,X}}\right)^{\frac{1-p_{n,X}}{p_{n,X}}} = \left(1+\frac{p_{n,X}}{1-p_{n,X}}\right)^{\frac{1-p_{n,X}}{p_{n,X}}} \geq 2, \qquad (\because\ p_{n,X} < 1/2)$$

so the conclusion follows. □

2.A.3 Proof of Theorem 2.7

Proof. Let U be a uniform random variable over the unit ball of $\mathbb{R}^d$ which is independent of X. Let also $U_1, U_2 \dots$ be independent copies of U, which is independent from $X_1, X_2, \dots$. We shall prove that $\lim_{\varepsilon \searrow 0} p_{n,X+\varepsilon U} = p_{n,X}$ for each n. Note that the distribution of $X + \varepsilon U$ has the probability density function

$$f(x) = \frac{1}{V\varepsilon^d}\mathbb{P}(\|X - x\|_2 \leq \varepsilon),$$

where V denotes the volume of the unit ball. Therefore, once we establish the limit $\lim_{\varepsilon \searrow 0} p_{n,X+\varepsilon U} = p_{n,X}$ the statement of the theorem is clear.

From $p_{d,X} = 0$, we know that

$$q_X(\delta) := \mathbb{P}\left(\inf_{y \in \operatorname{conv}\{X_i\}_{i=1}^d} \|y\| \leq \delta\right) \to 0, \qquad \delta \searrow 0. \tag{2.10}$$

For each $n \geq d+1$, consider the event $A_n := \{0 \in \operatorname{conv}\{X_1, \dots, X_n\}\}$. If the closed ε-ball centered at 0 is included in $\operatorname{conv}\{X_1, \dots, X_n\}$, then 0 is also contained in $\operatorname{conv}\{X_i + \varepsilon U_i\}_{i=1}^n$ as $\|\varepsilon U_i\| \leq \varepsilon$ for all i (more precisely, we can prove this by using the separating hyperplane theorem). Therefore, by considering the facets of the convex hull, we have

$$\mathbb{P}\left(A_n \cap \bigcap_{\substack{J \subset \{1,\dots,n\} \\ |J| = d}} \left\{\inf_{y \in \operatorname{conv}\{X_i\}_{i \in J}} \|y\| \geq \varepsilon\right\}\right) \leq \mathbb{P}(0 \in \operatorname{conv}\{X_i + \varepsilon U_i\}_{i=1}^n) = p_{n,X+\varepsilon U}.$$

By using (2.10), we have

$$\begin{aligned} p_{n,X+\varepsilon U} &\geq \mathbb{P}(A_n) - \mathbb{P}\left(\bigcup_{\substack{J\subset\{1,\ldots,n\}\\ |J|=d}} \left\{ \inf_{y\in\operatorname{conv}\{X_i\}_{i\in J}} \|y\| < \varepsilon \right\} \right) \\ &\geq p_{n,X} - \binom{n}{d} q_X(\varepsilon) \to p_{n,X} \qquad (\varepsilon \searrow 0), \end{aligned}$$

and so we obtain $\liminf_{\varepsilon\searrow 0} p_{n,X+\varepsilon U} \geq p_{n,X}$.

On the other hand, if we have $0 \in \operatorname{conv}\{X_i+\varepsilon U_i\}_{i=1}^n$ and $0 \notin \operatorname{conv}\{X_i\}_{i=1}^n$ at the same time, then there exists $J \subset \{1,\ldots,n\}$ such that $|J| = d$ and $\inf_{y\in\operatorname{conv}\{X_i\}_{i\in J}} \|y\| \leq \varepsilon$. Indeed, we can write 0 as a convex combination $\sum_{i=1}^n \lambda_i(X_i + \varepsilon U_i) = 0$, so

$$\left\| \sum_{i=1}^n \lambda_i X_i \right\| = \left\| \varepsilon \sum_{i=1}^n \lambda_i U_i \right\| \leq \varepsilon \sum_{i=1}^n \lambda_i \|U_i\| \leq \varepsilon.$$

As $0 \notin \operatorname{conv}\{X_i\}_{i=1}^n$, there is a facet within ε-distance from 0. Therefore, we obtain

$$\mathbb{P}(0 \in \operatorname{conv}\{X_i + \varepsilon U_i\}_{i=1}^n) \leq \mathbb{P}\left(A_n \cup \bigcup_{\substack{J\subset\{1,\ldots,n\}\\ |J|=d}} \left\{ \inf_{y\in\operatorname{conv}\{X_i\}_{i\in J}} \|y\| \leq \varepsilon \right\} \right),$$

and similarly, it follows that

$$p_{n,X+\varepsilon U} \leq p_{n,X} + \binom{n}{d} q_X(\varepsilon) \quad \textit{and} \quad \limsup_{\varepsilon\searrow 0} p_{n,X+\varepsilon U} \leq p_{n,X}.$$

Thus we finally obtain $\lim_{\varepsilon\searrow 0} p_{n,X+\varepsilon U} = p_{n,X}$. □

2.A.4 Proof of Proposition 2.8

Proof. For μ_M, it suffices to prove that with probability one there are no $J \subset \{1,\ldots,M\}$ with $|J| = d$ such that $0 \in \operatorname{conv}\{Y_i\}_{i\in J}$. This readily follows from the absolute continuity of the original measure μ. The extension to the case μ satisfies only $p_{d,\mu} = 0$ is immediate.

For the centered version $\widetilde{\mu}_M$, what to prove is that with probability one there are no $J \subset \{1,\ldots,M\}$ with $|J| = d$ such that $\frac{1}{M}\sum_{i=1}^M Y_j \in \operatorname{conv}\{Y_i\}_{i\in J}$. Suppose this occurs for some J. Then, we have that $\frac{1}{M-d}\sum_{i\neq J} Y_i$ is on the affine hull of $\{Y_i\}_{i\in J}$. However, as $\{Y_i\}_{i\notin J}$ is independent from $\{Y_i\}_{i\in J}$ for a fixed J, this probability is zero again from the absolute continuity of μ. Therefore, we have the desired conclusion. □

2.A.5 Proof of Proposition 2.10

Proof. If $\operatorname{dist}(0, \operatorname{conv}\{X_i\}_{i=1}^n) \le \varepsilon$, there exists a point $x \in \operatorname{conv}\{X_i\}_{i=1}^n$ with $\|x\| \le \varepsilon$. Then, for each $c \in \mathbb{R}^d$ with $\|c\| = 1$, we have $\langle c, x\rangle \le \varepsilon$ and so $\langle c, X_i\rangle \le \varepsilon$ for at least one $i \in \{1, \dots, n\}$. Hence we have a uniform evaluation

$$\begin{aligned}\mathbb{P}(\langle c, X\rangle \le \varepsilon) &= \frac{1}{n}\sum_{i=1}^n \mathbb{P}(\langle c, X_i\rangle \le \varepsilon)\\ &\ge \frac{1}{n}\mathbb{P}\left(\bigcup_{i=1}^n \{\langle c, X_i\rangle \le \varepsilon\}\right) \ge \frac{1}{n}\mathbb{P}(\operatorname{dist}(0, \operatorname{conv}\{X_i\}_{i=1}^n) \le \varepsilon)\,,\end{aligned}$$

and the first assertion follows.

For the latter, if α_X^ε is positive, we have $p_{n,X}^\varepsilon > 0$ for a sufficiently large n from Proposition 2.9. Finally, Carathéodory's theorem yields the positivity for all $n \ge d + 1$. □

2.A.6 Proof of Lemma 2.11

Proof. Note first that $g_{d,n}(\alpha)$ is non-increasing with respect to $\alpha \in [0, 1]$. Let $\widetilde{X}$ be a d-dimensional random vector such that $\|X - \widetilde{X}\| \le \delta$ for some $\delta > 0$. Then, for an arbitrary $c \in \mathbb{R}^d$ with $\|c\| = 1$, we have

$$\langle c, \widetilde{X}\rangle \le \langle c, X\rangle + \delta,$$

so $\mathbb{P}(\langle c, X\rangle \le \varepsilon) \le \mathbb{P}(\langle c, \widetilde{X}\rangle \le \varepsilon + \delta)$. Hence we have $\alpha_X^\varepsilon \le \alpha_{\widetilde{X}}^{\varepsilon+\delta}$.

Consider generating i.i.d. pairs $(X_1, \widetilde{X}_1), \dots, (X_n, \widetilde{X}_n)$ that are copies of $(X, \widetilde{X})$. Then, for each $x \in \operatorname{conv}\{X_i\}_{i=1}^n$, there is a convex combination such that $x = \sum_{i=1}^n \lambda_i X_i$ with $\lambda_i \ge 0$ and $\sum_{i=1}^n \lambda_i = 1$. Then, we have

$$\left\|x - \sum_{i=1}^n \lambda_i \widetilde{X}_i\right\| \le \sum_{i=1}^n \lambda_i \|X_i - \widetilde{X}_i\| \le \delta.$$

It means that $\inf_{y\in\mathrm{conv}\{\widetilde{X}_i\}_{i=1}^n}\|x-y\|\le\delta$ holds for every $x\in\mathrm{conv}\{X_i\}_{i=1}^n$, and we can deduce that $p_{n,X}^{\varepsilon+2\delta}\ge p_{n,\widetilde{X}}^{\varepsilon+\delta}$ holds.

In particular, we can choose $\widetilde{X}$ having density, so that we have $1-p_{n,X}^{\varepsilon+\delta}\le g_{d,n}(\alpha_{\widetilde{X}}^{\varepsilon+\delta})$. Therefore, from the monotonicity of $g_{d,n}$, we have

$$1-p_{n,X}^{\varepsilon+2\delta}\le 1-p_{n,\widetilde{X}}^{\varepsilon+\delta}\le g_{d,n}(\alpha_{\widetilde{X}}^{\varepsilon+\delta})\le g_{d,n}(\alpha_X^{\varepsilon}).$$

As $\delta>0$ can be taken arbitrarily, we finally obtain

$$1-p_{n,X}^{\varepsilon}\le g_{d,n}(\alpha_X^{\varepsilon})$$

by letting $\delta\to 0$. The δ-relaxation technique used in this proof is a big advantage of introducing $p_{n,X}^{\varepsilon}$ extending $p_{n,X}$. □

2.A.7 Proof of Theorem 2.12

First, from Lemma 2.11, we obtain the following general bound.

Proposition 2.24. *Let X be an arbitrary d-dimensional random vector. Then, for each $\varepsilon\ge 0$ and positive integer $n\ge d/\alpha_X^{\varepsilon}$, we have*

$$1-p_{n,X}^{\varepsilon}\le\left(\frac{n\alpha_X^{\varepsilon}}{d}\exp\left\{\left(\frac{1}{\alpha_X^{\varepsilon}}\log\frac{1}{1-\alpha_X^{\varepsilon}}\right)\left(1+\alpha_X^{\varepsilon}-\frac{n\alpha_X^{\varepsilon}}{d}\right)\right\}\right)^d.$$

Proof. From Lemma 2.11, it suffices to prove that

$$g_{d,n}(\alpha)\le\left(\frac{n\alpha}{d}\exp\left\{\left(\frac{1}{\alpha}\log\frac{1}{1-\alpha}\right)\left(1+\alpha-\frac{n\alpha}{d}\right)\right\}\right)^d \tag{2.11}$$

holds for each $\alpha\in(0,1)$ and $n\ge d/\alpha$. From the definition of $g_{d,n}$ (see (2.5)), if we set $n_0:=\lceil d/\alpha\rceil$, then we have

$$\begin{aligned}g_{d,n}(\alpha)&\le\frac{n(n-1)\cdots n_0}{(n-d)(n-d-1)\cdots(n_0-d)}(1-\alpha)^{n-n_0+1}g_{d,n_0-1}(\alpha)\\&\le\frac{n(n-1)\cdots(n-d+1)}{(n_0-1)(n_0-2)\cdots(n_0-d)}(1-\alpha)^{n-n_0+1}\\&\le\left(\frac{n}{n_0-d}\right)^d(1-\alpha)^{n-n_0+1}.\end{aligned}$$

As we know $d/\alpha \leq n_0 < d/\alpha + 1$ by definition, we have

$$g_{d,n}(\alpha) \leq \left(\frac{n}{d/\alpha - d}\right)^d (1-\alpha)^{n-\frac{d}{\alpha}} = \left(\frac{n\alpha}{d}\right)^d (1-\alpha)^{n-\frac{d}{\alpha}-d}.$$

This is indeed the desired inequality (2.11). □

Remark 2.8. As $\frac{1}{\alpha}\log\frac{1}{1-\alpha} \geq 1$ holds on $(0,1)$ for $n \geq \frac{(1+\alpha)d}{\alpha}$, the bound (2.11) yields a looser but more understandable variant

$$g_{d,n}(\alpha) \leq \left(\frac{n\alpha}{d}\exp\left(1+\alpha-\frac{n\alpha}{d}\right)\right)^d.$$

Note that we have a trivial lower bound of $1 - p_{n,X}^{\varepsilon} \geq (1-\alpha_X^{\varepsilon})^n$, which is proven by fixing a separating hyperplane between the origin and sample points.

By calculating the bound in Proposition 2.24 for a specific choice of $n = \lceil 3d/\alpha \rceil$, we can prove Theorem 2.12 as follows.

Proof of Theorem 2.12. From Proposition 2.24, it suffices to prove

$$3\exp\left\{\left(\frac{1}{\alpha}\log\frac{1}{1-\alpha}\right)(\alpha-2)\right\} < \frac{1}{2} \tag{2.12}$$

for all $\alpha \in (0,1)$. If we let $f(x) = \frac{x-2}{x}\log\frac{1}{1-x}$ for $x \in (0,1)$, then we have

$$f'(x) = \frac{1}{x^2}\left(2\log\frac{1}{1-x} - \frac{x(2-x)}{1-x}\right) = \frac{1}{x^2}\left(2\log\frac{1}{1-x} + (1-x) - \frac{1}{1-x}\right).$$

If we set $t := \log\frac{1}{1-x}$, t takes positive reals and we have

$$2\log\frac{1}{1-x} + (1-x) - \frac{1}{1-x} = 2t + e^{-t} - e^t = 2(t - \sinh t) < 0.$$

Therefore, it suffices to consider the limit $\alpha \searrow 0$. In this limit, the left-hand side of (2.12) is equal to $3e^{-2}$, which is smaller than $1/2$ since $e > \sqrt{6}$ holds. □

2.A.8 Proof of Proposition 2.14

Proof. First, observe that $p_{n+1,X} - p_{n,X} = \mathbb{P}(0 \in \operatorname{conv}\{X_1, \ldots, X_{n+1}\} \setminus \operatorname{conv}\{X_1, \ldots, X_n\})$ for $n \geq d+1$ and independent copies $X_1, X_2, \ldots$ of X. Assume $0 \in \operatorname{conv}\{X_1, \ldots, X_{n+1}\} \setminus \operatorname{conv}\{X_1, \ldots, X_n\}$ holds and no $d+1$ points of $\{0, X_1, \ldots, X_{n+1}\}$ lie on the same hyperplane (the latter is satisfied almost surely as X is absolutely continuous). Then, there exists an expression such that

$$0 = \sum_{i=1}^{n+1} \lambda_i X_i, \quad \sum_{i=1}^{n+1} \lambda_i = 1, \quad \lambda_i \geq 0.$$

Here $0 < \lambda_{n+1} < 1$ must hold as $0 \notin \mathrm{conv}\{X_1, \ldots, X_n\}$ and $X_{n+1} \neq 0$. Therefore, we can rewrite

$$\frac{1}{1-\lambda_{n+1}} \sum_{i=1}^{n} \lambda_i X_i = -\frac{\lambda_{n+1}}{1-\lambda_{n+1}} X_{n+1}$$

and this left-hand side is a convex combination of $\{X_1, \ldots, X_n\}$. Therefore, the line ℓ passing through X_{n+1} and 0 intersects $\mathrm{conv}\{X_1, \ldots, X_n\}$ after 0 (if directed from X_{n+1} to 0). Also, ℓ never intersects $\mathrm{conv}\{X_1, \ldots, X_n\}$ before 0. Indeed, if we have $\lambda X_{n+1} \in \mathrm{conv}\{X_1, \ldots, X_n\}$ for some $\lambda > 0$, then

$$0 \in \mathrm{conv}\left\{\lambda X_{n+1}, -\frac{\lambda_{n+1}}{1-\lambda_{n+1}} X_{n+1}\right\} \subset \mathrm{conv}\{X_1, \ldots, X_n\}$$

holds and it contradicts the assumption.

Hence, we can define the first hitting point of ℓ and $\mathrm{conv}\{X_1, \ldots, X_n\}$ after 0. More formally, let P be the minimum-normed point in $\ell \cap \mathrm{conv}\{X_1, \ldots, X_n\}$. Then, by the general-position assumption, there exists a unique $J \subset \{1, \ldots, n\}$ with $|J| = d$ such that $P \in \mathrm{conv}\{X_i\}_{i \in J}$ (more strongly, P is in the relative interior of $\mathrm{conv}\{X_i\}_{i \in J}$). In other words, $\mathrm{conv}\{X_i\}_{i \in J}$ is the unique facet which intersects ℓ first. Then, there exists a unique normal vector c_J that defines the hyperplane supporting $\{X_i\}_{i \in J}$, i.e., $\langle c_J, X_i \rangle = 1$ for each $i \in J$. Since $\langle c_J, P \rangle = 1$ also holds, we have $\langle c_J, X_{n+1} \rangle < 0$. We can also prove $\langle c_J, X_i \rangle > 1$ for each $i \in \{1, \ldots, n\} \setminus J$. Indeed, if we have $\langle c_J, X_j \rangle < 1$ for some $j \in \{1, \ldots, n\} \setminus J$, then there are interior points of $\mathrm{conv}\{X_i\}_{i \in J \cup \{j\}}$ that belongs to ℓ and this contradicts the minimality of the norm of P.

Therefore, for a fixed $J \subset \{1, \ldots, n\}$ with $|J| = d$, the probability that $0 \in \mathrm{conv}\{X_1, \ldots, X_{n+1}\} \setminus \mathrm{conv}\{X_1, \ldots, X_n\}$ holds and $\mathrm{conv}\{X_i\}_{i \in J}$ becomes the first facet intersecting ℓ after 0 is, from the independence,

$$\mathbb{E}\left[\mathbb{P}\big(0 \in \operatorname{conv}\{X_i\}_{i\in J\cup\{n+1\}} \mid \{X_i\}_{i\in J}\big) \prod_{j\in\{1,\dots,n\}\setminus J} \mathbb{P}(\langle c_J, X_j\rangle > 1 \mid \{X_i\}_{i\in J})\right]$$
$$= \mathbb{E}\Big[\mathbb{P}\big(0 \in \operatorname{conv}\{X_i\}_{i\in J\cup\{n+1\}} \mid \{X_i\}_{i\in J}\big)\, \mathbb{P}(\langle c_J, X'\rangle > 1 \mid \{X_i\}_{i\in J})^{n-d}\Big],$$

where X' is a copy of X independent from $\{X_i\}_{i\geq 1}$. By symmetry, this J is chosen with equal probability given $0 \in \operatorname{conv}\{X_1,\dots,X_{n+1}\} \setminus \operatorname{conv}\{X_1,\dots,X_n\}$ (almost surely without overlapping). Hence, we obtain

$$\begin{aligned}
&p_{n+1,X} - p_{n,X}\\
&= \binom{n}{d}\mathbb{E}\Big[\mathbb{P}(0 \in \operatorname{conv}\{X_1,\dots,X_{d+1}\} \mid \{X_i\}_{i\in I})\, \mathbb{P}(\langle c_I, X'\rangle > 1 \mid \{X_i\}_{i\in I})^{n-d}\Big],
\end{aligned}$$

where $I = \{1,\dots,d\}$. Observe that this representation is still valid for $n = d$. From the definition of α_X, we have $\mathbb{P}(\langle c_I, X'\rangle > 1 \mid \{X_i\}_{i\in I}) \leq 1 - \alpha_X$, so finally obtain, for $n \geq d+1$,

$$\begin{aligned}
&p_{n+1,X} - p_{n,X}\\
&= \binom{n}{d}\mathbb{E}\Big[\mathbb{P}(0 \in \operatorname{conv}\{X_1,\dots,X_{d+1}\} \mid \{X_i\}_{i\in I})\, \mathbb{P}(\langle c_I, X'\rangle > 1 \mid \{X_i\}_{i\in I})^{n-d}\Big]\\
&\leq (1-\alpha_X)\binom{n}{d}\mathbb{E}\Big[\mathbb{P}(0 \in \operatorname{conv}\{X_1,\dots,X_{d+1}\} \mid \{X_i\}_{i\in I})\, \mathbb{P}(\langle c_I, X'\rangle > 1 \mid \{X_i\}_{i\in I})^{n-1-d}\Big]\\
&= (1-\alpha_X)\frac{\binom{n}{d}}{\binom{n-1}{d}}(p_{n,X} - p_{n-1,X})\\
&= \frac{n(1-\alpha_X)}{n-d}(p_{n,X} - p_{n-1,X}).
\end{aligned}$$

This is the desired inequality. □

2.A.9 Proof of Proposition 2.15

Proof. For a one-dimensional random variable Y with $\mathbb{E}[Y] = 0$, $\mathbb{E}[Y^2] = 1$ and $|Y| \leq B$, we have

$$B\mathbb{P}(Y \leq 0) \geq \mathbb{E}[-\min\{Y, 0\}] = \frac{1}{2}\mathbb{E}[|Y|]$$

and so

$$\mathbb{P}(Y \leq 0) \geq \frac{\mathbb{E}[|Y|]}{2B} \geq \frac{\mathbb{E}[|Y|^2]}{2B^2} = \frac{1}{2B^2}.$$

By observing this inequality for each $Y = c^\top V^{-1/2}X$ with $\|c\|_2 = 1$, we obtain the bound of α_X. The latter bound then follows from Theorem 2.13. □

2.A.10 Proof of Theorem 2.17

Proof. Let n be an integer satisfying

$$n \geq \frac{9}{4} \sup_{c\in\mathbb{R}^d, \|c\|_2=1} \mathbb{E}\left[\left|c^\top V^{-1/2}X\right|^3\right]^2.$$

Then, for an arbitrary $\|c\|_2 = 1$, from Theorem 2.16, we have

$$\begin{aligned}\mathbb{P}\left(\frac{c^\top V^{-1/2}(X_1 + \cdots + X_n)}{n} \leq 0\right) &= \mathbb{P}\left(\frac{c^\top V^{-1/2}(X_1 + \cdots + X_n)}{\sqrt{n}} \leq 0\right)\\ &\geq \frac{1}{2} - \frac{2}{3}\cdot 0.48 = \frac{9}{50},\end{aligned}$$

where $X_1, X_2, \ldots$ are independent copies of X. Hence $\alpha_{n^{-1}(X_1+\cdots+X_n)} \geq 9/50$ holds. Then we can use Theorem 2.13 to obtain

$$N_{n^{-1}(X_1+\cdots+X_n)} \leq \left\lceil \frac{50}{9}\cdot 3d \right\rceil \leq 17d.$$

Since $N_X \leq nN_{n^{-1}(X_1+\cdots+X_n)}$ holds, we have

$$N_X \leq 17d\left(1 + \frac{9}{4} \sup_{c\in\mathbb{R}^d, \|c\|_2=1} \mathbb{E}\left[\left|c^\top V^{-1/2}X\right|^3\right]^2\right),$$

which is the desired conclusion. □

2.A.11 Proof of Corollary 2.18

Proof. From Theorem 2.17, it suffices to prove

$$\mathbb{E}\left[\left|c^\top V^{-1/2}X\right|^3\right]^2 \leq \mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]^2,\ \mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^4\right]$$

for each unit vector $c \in \mathbb{R}^d$. The first bound is clear from

$$\left|c^\top V^{-1/2}X\right| \leq \|c\|_2 \left\|V^{-1/2}X\right\|_2 = \left\|V^{-1/2}X\right\|_2.$$

The second bound can also be derived as

$$\mathbb{E}\left[\left|c^\top V^{-1/2}X\right|^3\right]^2 \le \mathbb{E}\left[\left|c^\top V^{-1/2}X\right|^2\right]\mathbb{E}\left[\left|c^\top V^{-1/2}X\right|^4\right] = \mathbb{E}\left[\left|c^\top V^{-1/2}X\right|^4\right]$$
$$\le \mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^4\right],$$

where we have used the Cauchy–Schwarz in the first inequality. □

2.A.12 Proof of Proposition 2.19

Proof. We fix α and denote

$$K_\varepsilon = \{\theta \in \mathbb{R}^d \mid \alpha_X^\varepsilon(\theta) \ge \alpha\}.$$

Note that $K_0 = K^\alpha(X)$. Let $c \in \mathbb{R}^d$ satisfy $\|c\| = 1$. Define $t(c)$ by

$$t(c) := \inf\{t \in \mathbb{R} \mid \mathbb{P}(\langle c, X\rangle \le t) \ge \alpha\}. \tag{2.13}$$

If $t(c) = \infty$, i.e., the right-hand set is empty for some c, then each set K_ε is empty. $t(c) > -\infty$ is clear from $\alpha > 0$. Suppose $t(c) \in \mathbb{R}$ for all c. From the continuity of probability, the infimum can actually be replaced by minimum, so we have

$$\mathbb{P}(\langle c, X - \theta\rangle \le \varepsilon) \ge \alpha \quad \Longleftrightarrow \quad \langle c, \theta\rangle + \varepsilon \ge t(c)$$

for each $\theta \in \mathbb{R}^d$. Hence, if $\theta_0 \in K_0$ and $\|\theta - \theta_0\| \le \varepsilon$, then we have $\theta \in K_\varepsilon$, so we obtain the inclusion statement.

Let us prove that K_ε is compact and convex. Define $H_\varepsilon(c) := \{\theta \in \mathbb{R}^d \mid \langle c, \theta\rangle \ge t(c) - \varepsilon\}$ for each $c \in \mathbb{R}^d$ with $\|c\| = 1$. From (2.13), we have $K_\varepsilon = \bigcap_{\|c\|=1} H_\varepsilon(c)$. As $H_\varepsilon(c)$ is closed and convex, K_ε is also closed and convex. To prove compactness, we shall prove K_ε is bounded. As X is a random vector, there is an $R > 0$ such that $\mathbb{P}(\|X\| \ge R) < \alpha$. Then, for each $\theta \in \mathbb{R}^d$ satisfying $\|\theta\| \ge R + \varepsilon$, we have

$$\mathbb{P}\left(\left\langle -\frac{\theta}{\|\theta\|}, X - \theta\right\rangle \le \varepsilon\right) = \mathbb{P}\left(\left\langle -\frac{\theta}{\|\theta\|}, X\right\rangle \le \varepsilon - \|\theta\|\right) \le \mathbb{P}(\|X\| \ge R) < \alpha.$$

Therefore, we have $\|\theta\| < R + \varepsilon$ for each $\theta \in K_\varepsilon$ and so K_ε is bounded. □

2.A.13 Proof of Proposition 2.20

Proof. Consider the set

$$A^\alpha := \{s \in \mathbb{R}^d \mid \mathbb{P}(\langle s, X\rangle \geq 1) < \alpha\}.$$

Then, we clearly have $A^\alpha \subset \widetilde{K}^\alpha(X)$ and so $(A^\alpha)^\circ \supset \big(\widetilde{K}^\alpha(X)\big)^\circ$. We first prove that $(A^\alpha)^\circ = K^\alpha(X)$ actually holds. From the definition of a polar, $\theta \in (A^\alpha)^\circ$ if and only if

$$\mathbb{P}(\langle s, X\rangle \geq 1) < \alpha \quad\Longrightarrow\quad \langle s, \theta\rangle \leq 1$$

holds for each $s \in \mathbb{R}^d \setminus \{0\}$. If we represent $s = r^{-1}c$ by $r > 0$ and $c \in \mathbb{R}^d$ with $\|c\| = 1$, this is equivalent to

$$\mathbb{P}(\langle c, X\rangle \geq r) < \alpha \quad\Longrightarrow\quad \langle c, \theta\rangle \leq r \tag{2.14}$$

for each $r > 0$ and $\|c\| = 1$. As we have assumed that X is symmetric and $\alpha < 1/2$, (2.14) is still equivalent even if we allow r to take all reals.

We shall prove that, for a fixed c, (2.14) is equivalent to $\mathbb{P}(\langle c, X - \theta\rangle \geq 0) \geq \alpha$. Indeed, if

$$\mathbb{P}(\langle c, X - \theta\rangle \geq 0) = \mathbb{P}(\langle c, X\rangle \geq \langle c, \theta\rangle) < \alpha$$

holds, there exists a $\delta > 0$ such that $\mathbb{P}(\langle c, X\rangle \geq \langle c, \theta\rangle - \delta) < \alpha$. Then, we have the negation of (2.14) by letting $r = \langle c, \theta\rangle - \delta$. For the opposite direction, if we assume $\mathbb{P}(\langle c, X\rangle \geq \langle c, \theta\rangle) \geq \alpha$, we have $\mathbb{P}(\langle c, X\rangle \geq r) \geq \alpha$ for all $r < \langle c, \theta\rangle$ and so (2.14) is true. Therefore, we obtain $(A^\alpha)^\circ = K^\alpha(X)$.

For each $\beta \in (\alpha, 1/2)$, we clearly have $\widetilde{K}^\alpha(X) \subset A^\beta$. Therefore, we have

$$\bigcup_{\alpha<\beta<1/2} K^\beta(X) \subset \big(\widetilde{K}^\alpha(X)\big)^\circ \subset K^\alpha(X),$$

which is the desired assertion. □

2.A.14 Proof of Proposition 2.21

Proof. We can only consider the case K has full dimension, i.e., K has a nonempty interior. Then, the Minkowski functional of K (e.g., see Conway [33, IV.1.14])

$$|||x||| := \inf\{t \mid t \geq 0,\ x \in tK\}$$

defines a norm on $\mathbb{R}^d$ (note that all norms are equivalent on $\mathbb{R}^d$). For this norm, it is known that there is a finite subset $A \subset S$ such that $\min_{y\in A} |||x-y||| \le \varepsilon$ for all $x \in B$ and $|A| \le (1+2/\varepsilon)^d$ [138, Lemma 4.10]. It suffices to prove $(1-\varepsilon)K \subset \operatorname{conv} A$. Assume the contrary, i.e., let x_0 be a point such that $|||x||| \le 1-\varepsilon$ and $x_0 \notin \operatorname{conv} A$. Then, there exists a $(d-1)$-dimensional hyperplane $H \subset \mathbb{R}^d$ such that $x_0 \in H$ and all the points in A lie (strictly) on the same side as the origin with respect to H. Let $y \in \operatorname{argmin}_{x\in H} |||x|||$. Then, we have $|||y||| \le 1-\varepsilon$, and $z := |||y|||^{-1} y$ satisfies $\min_{x\in H} |||z-x||| \ge \varepsilon$. Hence, we have $\min_{x\in A} |||z-x||| > \varepsilon$ and it contradicts the assumption for A. □

2.A.15 Proof of Theorem 2.22

Proof. As $K^\alpha(X)$ is symmetric and convex, there is a set $A \subset K^\alpha(X)$ with cardinality at most $(1+2/\varepsilon)^d$ such that $(1-\varepsilon)K^\alpha(X) \subset \operatorname{conv} A$ from Proposition 2.21. We shall evaluate the probability of $A \subset \operatorname{conv}\{X_i\}_{i=1}^n$. As each point $\theta \in A$ satisfies $\alpha_X(\theta) \ge \alpha$, from Remark 2.8, we have

$$1 - p_{n,X}(\theta) \le \left(\frac{n\alpha}{d}\exp\left(1+\alpha-\frac{n\alpha}{d}\right)\right)^d \tag{2.15}$$

for each $\theta \in A$. Hence, it suffices to prove the right-hand side of (2.15) is bounded by $(1+2/\varepsilon)^{-d}\delta$. By taking the logarithm, it is equivalent to showing

$$\frac{n\alpha}{d} - \log\frac{n\alpha}{d} \ge 1 + \alpha + \frac{\log(1/\delta)}{d} + \log\left(1+\frac{2}{\varepsilon}\right).$$

Let us denote $x := n\alpha/d$. For $x \ge 12$, as $x/2 - \log x$ is increasing, we have

$$\frac{x}{2} - \log x \ge 6 - \log 6 \ge 2 + \log 3 \ge 1 + \alpha + \log 3$$

by a simple computation. Therefore, from $\log(1+2/\varepsilon) \le \log 3 + \log(1/\varepsilon)$ and the assumption for n, we obtain the inequality (2.15). □

2.A.16 Proof of Corollary 2.23

Proof. For $\alpha = (en/d)^{-\beta}$, we have

$$\frac{\alpha}{12d}n = \frac{1}{12e^\beta}\left(\frac{n}{d}\right)^{1-\beta},$$

so $n \geq 12d/\alpha$ is equivalent to $n \geq (12e^{\beta})^{1/(1-\beta)}d$. Hence, from Theorem 2.22, it suffices to determine how small δ can be taken so as to satisfy

$$n \geq \frac{2d}{\alpha}\left(\frac{\log(1/\delta)}{d} + \log 2\right).$$

As $n \geq 12d$ holds for all β, for $a := \frac{\log 2}{6} < 0.1$, we have $an \geq \frac{2d}{\alpha}\log 2$. Therefore, we can take δ as small as

$$\log(1/\delta) = \frac{\alpha}{2}(1-a)n = \frac{1-a}{2}e^{-\beta}n^{1-\beta}d^{\beta}.$$

Therefore, we can take $c = \frac{1-a}{2} > 0.45$ as desired. □

2.B Bounds of N_X via multivariate Berry–Esseen theorem

In this section, we provide two different estimates of N_X. Although we can prove that the first bound (Section 2.B.2) is strictly stronger than the second one (Section 2.B.3), we also give the proof of the second as there seems to be more room for improvement in the second approach than in the first.

The following first bound is the one mentioned in (2.7). The proof is given in Section 2.B.2.

Theorem 2.25. *Let X be an $\mathbb{R}^d$-valued random vector which is centered and of nonsingular covariance matrix V. Then,*

$$N_X \leq 8d\left(1 + 36d^2(42d^{1/4}+16)^2\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]^2\right)$$

holds.

Note that

$$\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]^2 \geq \mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^2\right]^3 = d^3$$

holds so we can ignore the $\mathcal{O}(d)$ term. In the case $\sup\left\|V^{-1/2}X\right\|_2 < \infty$, we have

$$\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]^2 \leq \mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^2 \sup\left\|V^{-1/2}X\right\|_2\right]^2 = d^2\sup\left\|V^{-1/2}X\right\|_2^2.$$

Therefore, the following proposition, which only states $N_X = \widetilde{\mathcal{O}}\Bigl(d^{15/2}\sup\left\|V^{-1/2}X\right\|_2^2\Bigr)$, is weaker than Theorem 2.25. However, the approach of proofs is different and there seems to remain some room for improvement in the proof of Proposition 2.26, so we give the proof in Section 2.B.3.

Proposition 2.26. *Let X be an $\mathbb{R}^d$-valued random vector which is centered, bounded and of nonsingular covariance matrix V. Then, for all n satisfying*

$$\frac{n}{(1+\log n)^2} \le 2^{16}100d^{13/2}\sup\left\|V^{-1/2}X\right\|_2^2,$$

$N_X \le 6dn$ holds.

2.B.1 Multivariate Berry–Esseen bounds

Before proceeding to the evaluation of N_X, we briefly review multivariate Berry–Esseen type theorems. The following theorem should be the best known bound with explicit constants and dependence with respect to the dimension.

Theorem 2.27 ([140])**.** *Let $Y_1, \ldots, Y_n$ be i.i.d. D-dimensional independent random vectors with mean zero and covariance I_D. For any convex measurable set $A \subset \mathbb{R}^D$, it holds*

$$\left|\mathbb{P}\biggl(\frac{Y_1+\cdots+Y_n}{\sqrt{n}} \in A\biggr) - \mathbb{P}(Z\in A)\right| \le \frac{(42D^{1/4}+16)\mathbb{E}[\|Y_1\|_2^3]}{\sqrt{n}},$$

where Z is a D-dimensional standard Gaussian.

Note that the original statement is not limited to the i.i.d. case. However, similarly to the other existing Berry–Esseen type bounds, Theorem 2.27 only gives information about *convex* measurable sets. Thus we cannot use this result directly. However, Section 2.B.2 gives a creative use of Theorem 2.27.

Unlike the usual Berry–Esseen results, the next theorem can be used for non-convex cases with reasonable dependence on dimension. We denote by $\mathcal{W}_2(\mu,\nu)$ the Wasserstein-2 distribution between two probability measures μ and ν on the same domain. This is defined formally as

$$\mathcal{W}_2(\mu,\nu) := \inf_{Y\sim\mu, Z\sim\nu}\mathbb{E}\bigl[\|Y-Z\|_2^2\bigr],$$

where the infimum is taken for all the joint distribution (Y, Z) with the marginal satisfying $Y \sim \mu$ and $Z \sim \nu$. Although it is an abuse of notation, we also write $\mathcal{W}_2(Y, Z)$ to represent $\mathcal{W}_2(\mu, \nu)$ when $Y \sim \mu$ and $Z \sim \nu$ for some random variables Y and Z.

Theorem 2.28 ([182]). *Let $Y_1, \dots, Y_n$ be D-dimensional independent random vectors with mean zero, covariance Σ, and $\|Y_i\|_2 \le B$ almost surely for each i. If we let Z be a Gaussian with covariance Σ, then we have*

$$\mathcal{W}_2\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}, Z\right) \le \frac{5\sqrt{D}B(1 + \log n)}{\sqrt{n}}.$$

For a set $A \subset \mathbb{R}^D$ and an $\varepsilon > 0$, define

$$A^\varepsilon := \left\{ x \in \mathbb{R}^D \,\middle|\, \inf_{y \in A} \|x - y\|_2 \le \varepsilon \right\}, \qquad A^{-\varepsilon} := \left\{ x \in \mathbb{R}^D \,\middle|\, \inf_{y \in A^c} \|x - y\|_2 \ge \varepsilon \right\}.$$

By combining the following assertion with Theorem 2.28, we derive another bound of N_X in Section 2.B.3.

Proposition 2.29. *Let Y, Z be D-dimensional random vectors. Then, for any measurable set $A \subset \mathbb{R}^d$ and any $\varepsilon > 0$, the following estimates hold:*

$$\mathbb{P}(Y \in A) \le \mathbb{P}(Z \in A^\varepsilon) + \frac{\mathcal{W}_2(Y, Z)^2}{\varepsilon^2},$$
$$\mathbb{P}(Y \in A) \ge \mathbb{P}\big(Z \in A^{-\varepsilon}\big) - \frac{\mathcal{W}_2(Y, Z)^2}{\varepsilon^2}.$$

Proof. This proof is essentially the same as the argument given in the proof of [182, Proposition 1.4]. Let (Y', Z') be an arbitrary couple of random variables such that $Y' \sim Y$ and $Z' \sim Z$. Then, we have

$$\begin{aligned}\mathbb{P}(Y' \in A) &= \mathbb{P}(\|Y' - Z'\|_2 < \varepsilon,\ Y' \in A) + \mathbb{P}(\|Y' - Z'\|_2 \ge \varepsilon,\ Y \in A)\\ &\le \mathbb{P}(Z' \in A^\varepsilon) + \mathbb{P}(\|Y' - Z'\|_2 \ge \varepsilon)\\ &\le \mathbb{P}(Z' \in A^\varepsilon) + \frac{1}{\varepsilon^2}\mathbb{E}\big[\|Y' - Z'\|_2^2\big]. \qquad \text{(by Chebyshev's inequality)}\end{aligned}$$

By taking the infimum of the right-hand side with respect to all the possible couples (Y', Z'), we obtain the former result. The latter can also be derived by evaluating

$$\begin{aligned}\mathbb{P}\big(Z' \in A^{-\varepsilon}\big) &= \mathbb{P}\big(\|Y' - Z'\|_2 < \varepsilon,\ Z' \in A^{-\varepsilon}\big) + \mathbb{P}\big(\|Y' - Z'\|_2 \ge \varepsilon,\ Z \in A^{-\varepsilon}\big)\\ &\le \mathbb{P}(Y' \in A) + \mathbb{P}(\|Y' - Z'\|_2 \ge \varepsilon)\\ &\le \mathbb{P}(Y' \in A) + \frac{1}{\varepsilon^2}\mathbb{E}\big[\|Y' - Z'\|_2^2\big]\end{aligned}$$

and again taking the infimum. □

2.B.2 The first bound

In this section, we prove Theorem 2.25. We shall set $D = d$ and make use of Theorem 2.27.

First, fix a set $S \subset \mathbb{R}^d$ and consider the set $C(S) := \{x \in \mathbb{R}^d \mid 0 \in \operatorname{conv}(S \cup \{x\})\}$. We can prove this set is convex for any S. Indeed, if $0 \in \operatorname{conv} S$, then clearly $C(S) = \mathbb{R}^d$. Otherwise, $x \in C(S)$ is equivalent to the existence of some $k \geq 0$ and $x_1, \dots, x_k \in S$, $\lambda > 0$, $\lambda_1, \dots, \lambda_k \geq 0$ such that

$$\lambda + \lambda_1 + \cdots + \lambda_k = 1, \qquad \lambda x + \lambda_1 x_1 + \cdots + \lambda_k x_k = 0.$$

Here, $\lambda > 0$ comes from the assumption $0 \notin \operatorname{conv} S$. This occurs if and only if x is contained in the negative cone of S, i.e., $C(S) = \{\sum_{i=1}^k \widetilde{\lambda}_i x_i \mid k \geq 0,\ \widetilde{\lambda}_i \leq 0,\ x_i \in S\}$. In both cases, $C(S)$ is convex, so S_0 is always convex (and of course measurable).

Let X be an $\mathbb{R}^d$-valued random vector with mean 0 and nonsingular covariance V. Suppose $\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right] < \infty$. Let $X_1, X_2, \dots$ be independent copies of X, and for a fixed positive integer n, define

$$W_i := \frac{V^{-1/2}X_{(i-1)n+1} + \cdots + V^{-1/2}X_{in}}{\sqrt{n}}$$

for $i = 1, \dots, 2d$. We also let $Z_1, \dots, Z_{2d}$ be independent d-dimensional standard Gaussian which is also independent from $X_1, X_2, \dots$. Then, by using Theorem 2.27 and the above-mentioned convexity of $C(S)$, we have

$$\begin{aligned}\mathbb{P}(0 \in \{W_1, \dots, W_{2d}\}) &= \mathbb{P}(W_1 \in C(\{W_2, \dots, W_{2d}\})) \\ &\geq \mathbb{P}(Z_1 \in C(\{W_2, \dots, W_{2d}\})) - \frac{(42d^{1/4} + 16)\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]}{\sqrt{n}} \\ &= \mathbb{P}(0 \in \operatorname{conv}\{Z_1, W_2, \dots, W_{2d}\}) - \frac{(42d^{1/4} + 16)\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]}{\sqrt{n}}.\end{aligned}$$

By repeating similar evaluations, we obtain

$$\begin{aligned}
&\mathbb{P}(0 \in \mathrm{conv}\{W_1, \dots, W_{2d}\}) \\
&\geq \mathbb{P}(0 \in \mathrm{conv}\{Z_1, W_2, \dots, W_{2d}\}) - \frac{(42d^{1/4} + 16)\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]}{\sqrt{n}} \\
&\geq \mathbb{P}(0 \in \mathrm{conv}\{Z_1, Z_2, W_3, \dots, W_{2d}\}) - \frac{2(42d^{1/4} + 16)\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]}{\sqrt{n}} \\
&\vdots \\
&\geq \mathbb{P}(0 \in \mathrm{conv}\{Z_1, \dots, Z_i, W_{i+1}, \dots, W_{2d}\}) - \frac{i(42d^{1/4} + 16)\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]}{\sqrt{n}} \\
&\vdots \\
&\geq \mathbb{P}(0 \in \mathrm{conv}\{Z_1, \dots, Z_{2d}\}) - \frac{2d(42d^{1/4} + 16)\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]}{\sqrt{n}} \\
&= \frac{1}{2} - \frac{2d(42d^{1/4} + 16)\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]}{\sqrt{n}}.
\end{aligned}$$

Therefore, by letting

$$n = \left\lceil 36d^2(42d^{1/4} + 16)^2\mathbb{E}\left[\left\|V^{-1/2}X\right\|_2^3\right]^2 \right\rceil,$$

we have $\mathbb{P}(0 \in \mathrm{conv}\{X_1, \dots, X_{2dn}\}) \geq 1/6$. Since $(1-1/6)^4 < 1/2$ holds, we finally obtain $N_X \leq 8dn$.

2.B.3 The second bound

In this section, we provide a proof of Section 2.26 in a different manner from the one given in the previous section. We set $D = 2d^2$ and define $A_d \subset \mathbb{R}^D$ as follows:

$$A_d := \{x = (x_1, \dots, x_{2d}) \in (\mathbb{R}^d)^{2d} \simeq \mathbb{R}^D \mid 0 \in \mathrm{conv}\{x_1, \dots, x_{2d}\} \subset \mathbb{R}^d\}.$$

Then, it suffices to find a suitable upper bound of $\mathbb{P}\big(Z \in A_d \setminus A_d^{-\varepsilon}\big)$ for a D-dimensional standard Gaussian Z for our purpose. For an $\varepsilon > 0$, $B_{d,\varepsilon} := A_d \setminus A_d^{-\varepsilon}$ can be explicitly written as

$$B_{d,\varepsilon} = \left\{ x = (x_1, \dots, x_{2d}) \in \mathbb{R}^D \;\middle|\; \begin{array}{c} 0 \in \mathrm{conv}\{x_i\}_{i=1}^{2d}, \\ \exists \widetilde{x} = (\widetilde{x}_i)_{i=1}^{2d} \in \mathbb{R}^D \text{ s.t.} \|x - \widetilde{x}\|_2 < \varepsilon,\ 0 \notin \mathrm{conv}\{\widetilde{x}_i\}_{i=1}^{2d} \end{array} \right\}. \tag{2.16}$$

For a (finite) set $S = \{v_1, \dots, v_j\} \subset \mathbb{R}^d$, define the negative box $N(S) \subset \mathbb{R}^d$ by

$$N(S) := \{a_1 v_1 + \cdots + a_j v_j \mid a_i \in [-1, 0]\}.$$

$N(S)$ is obviously a convex set.

Lemma 2.30. *For an arbitrary $x = (x_1, \dots, x_{2d}) \in B_{d,\varepsilon}$, there exists an index $k \in \{1, \dots, 2d\}$ such that $x_k \in N(\{x_i \mid i \neq k\}) \setminus N(\{x_i \mid i \neq k\})^{-\varepsilon\sqrt{2d}}$.*

Proof. As $0 \in \operatorname{conv}\{x_i\}_{i=1}^{2d}$, there exist nonnegative weights $\lambda_1, \dots, \lambda_{2d}$ such that $\lambda_1 x_1 + \cdots + \lambda_{2d} x_{2d} = 0$ with the total weight one. Let k be an index such that w_k is the maximum weight. Then, λ_k is clearly positive and we have $x_k = \sum_{i \neq k} -\frac{\lambda_i}{\lambda_k} x_i$. Therefore, we obtain $x_k \in N(\{x_i \mid i \neq k\})$.

By (2.16), there exists an $\widetilde{x} = (\widetilde{x}_i)_{i=1}^{2d} \in \mathbb{R}^D$ such that $\sum_{i=1}^{2d} \|x_i - \widetilde{x}_i\|_2^2 < \varepsilon^2$ and $0 \notin \operatorname{conv}\{\widetilde{x}_i\}_{i=1}^{2d}$. We can prove that $\widetilde{x}_k \notin N(\{\widetilde{x}_i \mid i \neq k\})$. Indeed, if we can write $\widetilde{x}_k = -\sum_{i \neq k} a_i \widetilde{x}_i$ with $a_i \in [0, 1]$, then

$$\left(1 + \sum_{i \neq k} a_i\right)^{-1} \left(\widetilde{x}_k + \sum_{i \neq k} a_i \widetilde{x}_i\right) = 0$$

is a convex combination and it contradicts the assumption $0 \notin \operatorname{conv}\{\widetilde{x}_i\}_{i=1}^{2d}$. Therefore, we can take a unit vector $c \in \mathbb{R}^d$ such that

$$c^\top \widetilde{x}_k > \max\{c^\top y \mid y \in N(\{\widetilde{x}_i \mid i \neq k\})\}. \tag{2.17}$$

Let us assume the closed ball with center x_k and radius δ is included in $N(\{x_i \mid i \neq k\})$ for a $\delta > 0$. Then, if $\delta > \|x_k - \widetilde{x}_k\|_2$, the closed ball with center $\widetilde{x}_k$ and radius $\delta' := \delta - \|x_k - \widetilde{x}_k\|_2$ is included in $N(\{x_i \mid i \neq k\})$. In particular, we have some coefficients $a_i \in [-1, 0]$ such that $\widetilde{x}_k + \delta' c = \sum_{i \neq k} a_i x_i$. By the inequality (2.17), we have

$$c^\top \widetilde{x}_k > c^\top \sum_{i \neq k} a_i \widetilde{x}_i = c^\top \left(\widetilde{x}_k + \delta' c + \sum_{i \neq k} a_i (\widetilde{x}_i - x_i)\right),$$

so by arranging

$$\delta' < \sum_{i \neq k} a_i c^\top (x_i - \widetilde{x}_i) \leq \sum_{i \neq k} \|x_i - \widetilde{x}_i\|_2.$$

Therefore, from the definition of δ', we obtain

$$\delta < \sum_{i=1}^{2d} \|x_i - \widetilde{x}_i\|_2 \le \left(2d \sum_{i=1}^{2d} \|x_i - \widetilde{x}_i\|_2\right)^{1/2} \le \varepsilon\sqrt{2d}$$

by Cauchy-Schwarz and the assumption. It immediately implies the desired assertion. □

Proposition 2.31. *$\mathbb{P}(Z \in B_{d,\varepsilon}) \le 8\sqrt{2}d^{7/4}\varepsilon$ holds.*

Proof. By Lemma 2.30, we have $B_{d,\varepsilon} \subset \bigcup_{k=1}^{2d}\{x \mid x_k \in N(\{x_i \mid i \ne k\}) \setminus N(\{x_i \mid i \ne k\})^{-\varepsilon\sqrt{2d}}\}$. Therefore, letting $Z = (Z_1, \dots, Z_{2d})$ be a standard Gaussian in $\mathbb{R}^D$ (where each Z_i is a independent standard Gaussian in $\mathbb{R}^d$), we can evaluate

$$\mathbb{P}(Z \in B_{d,\varepsilon}) \le \sum_{k=1}^{2d} \mathbb{P}\Big(Z_k \in N(\{Z_i \mid i \ne k\}) \setminus N(\{Z_i \mid i \ne k\})^{-\varepsilon\sqrt{2d}}\}\Big).$$

For each k, Z_k is independent from the random convex set $N(\{Z_i \mid i \ne k\})$. Therefore, we can use the result of Ball [8] to deduce

$$\mathbb{P}\Big(Z_k \in N(\{Z_i \mid i \ne k\}) \setminus N(\{Z_i \mid i \ne k\})^{-\varepsilon\sqrt{2d}}\}\Big) \le 4d^{1/4} \cdot \varepsilon\sqrt{2d}.$$

Therefore, we finally obtain

$$\mathbb{P}(Z \in B_{d,\varepsilon}) \le 2d \cdot 4d^{1/4} \cdot \varepsilon\sqrt{2d} = 8\sqrt{2}d^{7/4}\varepsilon.$$

□

By letting $\varepsilon = 2^{-13/2}d^{-7/4}$, we have $\mathbb{P}(Z \in B_{d,\varepsilon}) \le 1/8$. Under this value of ε, if we let n satisfy

$$\frac{n}{(1+\log n)^2} \ge \frac{8 \cdot 25DB^2}{\varepsilon^2} = 400d^2B^2 \cdot 2^{13}d^{7/2} = 2^{15}100B^2d^{11/2}, \tag{2.18}$$

for a constant B, then we have

$$\left(\frac{5\sqrt{D}B(1+\log n)}{\sqrt{n}}\right)^2 \le \frac{\varepsilon^2}{8}.$$

Now consider a bounded and centered $\mathbb{R}^d$-valued random vector X with $V = \mathbb{E}\big[XX^\top\big]$ nonsingular. Then $B' := \sup\big\|V^{-1/2}X\big\|_2$ is finite. Let $X_1, X_2, \dots$ be independent copies of X. Define $\mathbb{R}^D$-valued random vectors $Y_1, Y_2, \dots$ by $Y_i := (V^{-1/2}X_{(2i-1)d+1}, \dots, V^{-1/2}X_{2id})^\top$ for each i. Then, note that $\|Y_i\|_2 \le \sqrt{2d}B'$. By taking $B = \sqrt{2d}B'$ in (2.18), we have from Theorem 2.28 that (for $\varepsilon = 2^{-13/2}d^{-7/4}$)

$$\mathbb{P}(Z \in B_{d,\varepsilon}) \le \frac{1}{8}, \qquad \frac{1}{\varepsilon^2}\mathcal{W}_2\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}, Z\right) \le \frac{1}{8}.$$

From Proposition 2.29, we obtain

$$\mathbb{P}\biggl(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}} \in A_d\biggr) \ge \mathbb{P}(Z \in A_d) - \mathbb{P}(Z \in B_{d,\varepsilon}) - \frac{1}{\varepsilon^2}\mathcal{W}_2\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}, Z\right) \ge \frac{1}{4}.$$

Therefore, 0 is contained in the convex hull of $\{X_1, \dots, X_{2dn}\}$ with probability at least $1/4$. Since $(1 - 1/4)^3 < 1/2$, $N_X \le 6dn$ holds. Therefore, our proof of Proposition 2.26 is complete.

2.C Extreme examples

Before treating concrete examples, we prove a proposition which is useful for evaluating N_X.

Lemma 2.32. *For a random vector X and its independent copies $X_1, X_2, \dots$, define $\widetilde{N}_X$ as the minimum index n satisfying $0 \in \operatorname{conv}\{X_1, \dots, X_n\}$. Then, we have*

$$\frac{1}{2}\mathbb{E}\left[\widetilde{N}_X\right] \le N_X \le 2\mathbb{E}\left[\widetilde{N}_X\right].$$

Proof. From the definition of N_X, $\mathbb{P}(0 \in \{X_1, \dots, X_{N_X - 1}\}) < 1/2$ holds. Thus $\mathbb{P}\left(\widetilde{N}_X \ge N_X\right) \ge 1/2$, and so we obtain $\mathbb{E}\left[\widetilde{N}_X\right] \ge \frac{1}{2}N_X$.

For the other inequality, we use the evaluation $\mathbb{P}\left(\widetilde{N}_X \ge kN_X\right) \le 2^{-k}$ for each nonnegative integer k. As $\widetilde{N}_X$ is a nonnegative discrete random variable,

$$\mathbb{E}\left[\widetilde{N}_X\right] = \sum_{n=1}^{\infty} \mathbb{P}\left(\widetilde{N}_X \ge n\right) \le \sum_{k=0}^{\infty} N_X \mathbb{P}\left(\widetilde{N}_X \ge kN_X\right) \le 2N_X$$

holds. □

Note that all the examples given below satisfy $p_{d,X} = 0$. They are given as one of the worst-case examples for uniform estimates of N_X in Proposition 2.4 or Theorem 2.25. Let us start with the simplest extreme case.

Example 2.33. Let $d = 1$. For an $\varepsilon \in (0,1)$, let X be a random variable such that $\mathbb{P}(X = 1/\varepsilon) = \varepsilon$ and $\mathbb{P}(X = -1/(1-\varepsilon)) = 1-\varepsilon$. Then $\mathbb{E}[X] = 0$.

In this example, we can explicitly calculate $p_{n,X}$ as

$$p_{n,X} = 1 - \varepsilon^n - (1-\varepsilon)^n.$$

In particular, $p_{2,X} = 2\varepsilon - 2\varepsilon^2$. We have $\lim_{\varepsilon \searrow 0}(1-\varepsilon)^{1/2\varepsilon} = e^{-1/2} = 0.60\ldots$, so $p_{\lceil 1/2\varepsilon \rceil, X} < 1/2$ holds for a sufficiently small ε. For such an ε, we have

$$N_X \geq \frac{1}{2\varepsilon} = \frac{1-\varepsilon}{2} \frac{2}{p_{2,X}}, \tag{2.19}$$

and so $N_X \leq \frac{2}{p_{2,X}}$ in Proposition 2.4 is sharp up to constant.

For $\varepsilon \in (0,1/2)$, N_X can also be evaluated above as $N_X \leq 2\mathbb{E}\left[\widetilde{N}_X\right] \leq 2\left(\frac{1}{\varepsilon} + \frac{1}{(1-\varepsilon)}\right)$ by using Proposition 2.32. We also have $\alpha_X = \varepsilon$ for $\varepsilon \in (0,1/2)$, so

$$\inf_{X:1\text{-dimensional}} \alpha_X N_X \leq 2 + \frac{2\varepsilon}{1-\varepsilon} \to 2 \quad (\varepsilon \to 0).$$

As the variance is $V = \mathbb{E}[X^2] = \frac{1}{\varepsilon} + \frac{1}{1-\varepsilon} = \frac{1}{\varepsilon(1-\varepsilon)}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left|V^{-1/2}X\right|^3\right]^2 &= V^{-3}\left(\frac{1}{\varepsilon^2} + \frac{1}{(1-\varepsilon)^2}\right)^2 \\
&= \varepsilon^3(1-\varepsilon)^3\left(\frac{1}{\varepsilon^4} + \frac{2}{\varepsilon^2(1-\varepsilon)^2} + \frac{1}{(1-\varepsilon)^4}\right) \\
&= \frac{1}{\varepsilon} + \mathcal{O}(1)\,.
\end{aligned}$$

Therefore, from (2.19), we obtain

$$\sup\left\{\mathbb{E}\left[\left|V^{-1/2}X\right|^3\right]^{-2} N_X \;\middle|\; \begin{array}{c} X \text{ is 1-dimensional, } \mathbb{E}[X] = 0, \\ V = \mathbb{E}[X^2] \in (0,\infty),\ \mathbb{E}\left[\left|V^{-1/2}X\right|^3\right] < \infty \end{array}\right\} \geq \frac{1}{2},$$

which is what is mentioned in Remark 2.5 when $d = 1$.

The next example is a multi-dimensional version of the previous one.

Example 2.34. Let $d \geq 2$. Let $\{e_1, \ldots, e_d\} \subset \mathbb{R}^d$ be the standard basis of $\mathbb{R}^d$. Let us first consider, for an arbitrary $\varepsilon \in (0,1)$, a random vector X given by

$$X = Y\left(\sum_{i=1}^{d-1} Z^i e_i - \frac{1}{1-\varepsilon}e_d\right) + \frac{1}{\varepsilon}(1-Y)e_d,$$

where $\mathbb{P}(Y = 1) = 1 - \varepsilon$, $\mathbb{P}(Y = 0) = \varepsilon$ and $Z^1, \dots, Z^{d-1}$ are independent uniform random variables over $[-1, 1]$. (also independent from Y). Namely, X is $\varepsilon^{-1}e_d$ with probability ε and a $(d-1)$-dimensional uniform vector over a box on the hyperplane $\{x \in \mathbb{R}^d \mid e_d^\top x = -(1-\varepsilon)^{-1}\}$ otherwise. $\mathbb{E}[X] = 0$ also holds.

Let us estimate $p_{d+1,X}, p_{2d,X}$ and N_X for this X. To contain the origin in the convex hull, we have to observe at least one X_i with $Y = 0$. Therefore, for an $\varepsilon \ll 1/d$, we have

$$\begin{aligned}
p_{d+1,X} &= (d+1)\varepsilon(1-\varepsilon)^d 2^{-(d-1)} = \frac{d+1}{2^{d-1}}\varepsilon\left(1 + \mathcal{O}\big(d^2\varepsilon^2\big)\right) \\
p_{2d,X} &= \sum_{k=1}^{d} \binom{2d}{k} \varepsilon^k (1-\varepsilon)^{2d-k} p_{2d-k,X'} \\
&= 2d\varepsilon p_{2d-1,X'} + \mathcal{O}\big(d^2\varepsilon^2\big) = d\left(1 + \frac{1}{2^{2d-2}}\binom{2d-2}{d-1}\right)\varepsilon + \mathcal{O}\big(d^2\varepsilon^2\big) \\
&\geq d\left(1 + \frac{1}{2\sqrt{d-1}}\right)\varepsilon + \mathcal{O}\big(d^2\varepsilon^2\big),
\end{aligned}$$

where X' represents a $(d-1)$-dimensional uniform random vector over the box $[-1,1]^{d-1}$. We can see that $p_{2d,X} \gtrsim 2^{d-1} p_{d+1,X}$ holds for a small ε as Remark 2.2 suggests.

For the calculation of N_X, we can exploit Proposition 2.32. We first bound the expectation of $\widetilde{N}_X$. For independent copies $X_1, X_2, \dots$ of X, let N_1 be the minimum integer n satisfying $X_n = \varepsilon^{-1}e_d$. We also define N_2 as the minimum integer n satisfying $-(1-\varepsilon)^{-1}e_d \in \operatorname{conv}\{X_1, \dots, X_n\}$. Then, $\widetilde{N}_X = \max\{N_1, N_2\}$ holds. Thus we have $N_1 \leq \widetilde{N}_X \leq N_1 + N_2$. $\mathbb{E}[N_1] = 1/\varepsilon$ clearly holds. For N_2, we can evaluate (again using X') as

$$\mathbb{E}[N_2] = \frac{1}{1-\varepsilon}\mathbb{E}\left[\widetilde{N}_{X'}\right] \leq \frac{2N_{X'}}{1-\varepsilon} = \frac{4(d-1)}{1-\varepsilon},$$

where we have used Proposition 2.32 for the inequality. Therefore, from Proposition 2.32, we obtain

$$\frac{1}{2\varepsilon} \leq \frac{1}{2}\mathbb{E}\left[\widetilde{N}_X\right] \leq N_X \leq 2\mathbb{E}\left[\widetilde{N}_X\right] \leq \frac{2}{\varepsilon} + \frac{8(d-1)}{1-\varepsilon}. \tag{2.20}$$

We finally compare the naive general estimate $N_X \le \frac{n}{p_{n,X}}$ in Proposition 2.4 with this example. From (2.20), we have

$$\frac{N_X p_{2d,X}}{2d} \ge \frac{p_{2d,X}}{4d\varepsilon} \ge \frac{1}{4} + \frac{1}{8\sqrt{d-1}} + \mathcal{O}(d\varepsilon)\,.$$

Therefore, the evaluation $N_X \le \frac{2d}{p_{2d,X}}$ is sharp even for small $p_{2d,X}$ up to constant in the sense that we have

$$\lim_{\varepsilon\to 0} \sup_{\substack{X: d\text{-dimensional} \\ p_{2d,X}<\varepsilon}} \frac{N_X p_{2d,X}}{2d} \ge \frac{1}{4} + \frac{1}{8\sqrt{d-1}}.$$

Also in this example, we have $\alpha_X = \varepsilon$ for $\varepsilon \in (0, 1/3)$. Hence, combined with (2.20), we have

$$\alpha_X N_X \le \varepsilon\left(\frac{2}{\varepsilon} + \frac{8(d-1)}{1-\varepsilon}\right) = 2 + \frac{8(d-1)\varepsilon}{1-\varepsilon} \to 2 \quad (\varepsilon \to 0).$$

Therefore, we have $\inf_{X: d\text{-dim}} \alpha_X N_X \le 2$.

We next evaluate the value of $\mathbb{E}\Big[\big\|V^{-1/2}X\big\|_2^3\Big]$, where $V = (V^{ij})$ is the covariance matrix of X with respect to the basis $\{e_1, \ldots, e_d\}$. Then, for $(i,j) \in \{1, \ldots, d-1\}^2$, we obtain

$$V^{ij} = \mathbb{E}\big[Y^2 Z^i Z^j\big] = \mathbb{E}\big[Y^2\big]\,\mathbb{E}\big[Z^i Z^j\big] = \frac{1-\varepsilon}{2}\delta^{ij}, \qquad (\delta^{ij}\text{: Kronecker's delta})$$
$$V^{id} = \mathbb{E}\left[Y Z^i \left(-\frac{Y}{1-\varepsilon} + \frac{1-Y}{\varepsilon}\right)\right] = \mathbb{E}\big[Z^i\big]\,\mathbb{E}\left[Y\left(-\frac{Y}{1-\varepsilon} + \frac{1-Y}{\varepsilon}\right)\right] = 0$$

by using the independence of Y, $Z_1, \ldots, Z_{d-1}$. For the V^{dd}, we have

$$V^{dd} = \frac{1}{1-\varepsilon} + \frac{1}{\varepsilon} = \frac{1}{\varepsilon(1-\varepsilon)}.$$

Therefore, $V^{-1/2}X$ can be explicitly written as

$$V^{-1/2}X = Y\left(\sqrt{\frac{2}{1-\varepsilon}}\sum_{i=1}^{d-1} Z^i e_i - \sqrt{\frac{\varepsilon}{1-\varepsilon}}e_d\right) + \sqrt{\frac{1-\varepsilon}{\varepsilon}}(1-Y)e_d.$$

Thus we have

$$\big\|V^{-1/2}X\big\|_2^2 \le Y\frac{2(d-1)+\varepsilon}{1-\varepsilon} + (1-Y)\frac{1-\varepsilon}{\varepsilon},$$

and so

$$\mathbb{E}\Big[\big\|V^{-1/2}X\big\|_2^3\Big] \le \frac{(2(d-1)+\varepsilon)^{3/2}}{\sqrt{1-\varepsilon}} + \frac{(1-\varepsilon)^{3/2}}{\sqrt{\varepsilon}} \le 4d^{3/2} + \varepsilon^{-1/2}$$

holds when $0 < \varepsilon < 1/2$. By using (2.20), we obtain

$$\frac{N_X}{\mathbb{E}\Big[\|V^{-1/2}X\|_2^3\Big]^2} \ge \frac{1}{2\varepsilon(4d^{3/2}+\varepsilon^{-1/2})^2} = \frac{1}{2(4d^{3/2}\varepsilon^{1/2}+1)^2}.$$

Therefore, by taking $\varepsilon \to 0$, we finally obtain the estimate

$$\sup\left\{ \frac{N_X}{\mathbb{E}\Big[\|V^{-1/2}X\|_2^3\Big]^2} \;\middle|\; \begin{array}{c} X \text{ is } d\text{-dimensional, } \mathbb{E}[X]=0, \\ V=\mathbb{E}[X^2] \text{ is nonsingular, } \mathbb{E}\Big[\big\|V^{-1/2}X\big\|^3\Big] < \infty \end{array} \right\} \ge \frac{1}{2}$$

as mentioned in Remark 2.5.

Chapter 3

Hypercontractivity meets random convex hulls

Given a test function vector $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^D$ and a random variable x taking values in $\mathcal{X}$, we have analyzed the naive random sampling for constructing a cubature in the previous chapter. However, it is not fully sufficient to explain what happens in the more concrete interesting cases. Here we analyze the computational complexity of this approach when $\boldsymbol{\varphi}$ exhibits a graded structure by using so-called hypercontractivity. The resulting theorem covers not only the classical cubature case of multivariate polynomials but also integration on path-space and kernel quadrature for product measures.

3.1 Introduction

Let X be a random variable that takes values in a set $\mathcal{X}$, and $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ a linear, finite-dimensional space of functions satisfying $\mathbb{E}[|f(X)|] < \infty$ for each $f \in \mathcal{F}$. A cubature formula for $(X, \mathcal{F})$ is a finite set of points $(x_i) \subset \mathcal{X}$ and weights $(w_i) \subset \mathbb{R}$ such that

$$\mathbb{E}[f(X)] = \sum_{i=1}^{n} w_i f(x_i) \text{ for all } f \in \mathcal{F}. \tag{3.1}$$

This is a reformulation of cubature introduced in Section 1.1 from the viewpoint of function spaces. We also denote $\mu = \mathrm{Law}(X)$ and refer to $\hat{\mu} = \sum_{i=1}^{n} w_i \delta_{x_i}$ as the cubature measure for $(X, \mathcal{F})$. The existence of such a cubature formula that further satisfies $n \leq 1 + \dim \mathcal{F}$, $w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$ is guaranteed by Tchakaloff's theorem (Theorem 1.1). Arguably the most famous applications concern the case when $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and $\mathcal{F}$ is the linear space of polynomials up to a certain degree, that is $\mathcal{F}$ is spanned by monomials up to a certain degree. However, more recent applications include the case when $\mathcal{X}$ is a space of paths and $\mathcal{F}$ is spanned by iterated Ito–Stratonovich integrals [110], or kernel quadrature [84] (see also Chapter 4) where $\mathcal{X}$ is a set that carries a positive definite kernel and $\mathcal{F}$ is a subset of the associated reproducing kernel Hilbert space that is spanned by some "test functions" including eigenfunctions of the integral operator induced by this kernel.

Random convex hulls. If $\mathcal{F}$ is spanned by m functions $\varphi_1, \ldots, \varphi_m : \mathcal{X} \to \mathbb{R}$, then we can denote $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_m) : \mathcal{X} \to \mathbb{R}^m$ and see that (3.1) is equivalent to $\mathbb{E}[\boldsymbol{\varphi}(X)] = \sum_{i=1}^n w_i \boldsymbol{\varphi}(x_i)$. Given an $N(\gg n)$ i.i.d. sample $X_1, \ldots, X_N \sim X$, we want to know the probability of the event

$$\mathbb{E}[\boldsymbol{\varphi}(X)] \in \operatorname{conv}\{\boldsymbol{\varphi}(X_1), \ldots, \boldsymbol{\varphi}(X_N)\}, \tag{3.2}$$

under which we can construct the desired cubature satisfying (3.1) by using linear programming as explained in Section 1.1.

Empirically, this approach turns out to work well already for "reasonable" magnitudes of N [66] (also see the experiments in the next chapter). The aim of this chapter is to provide theoretical guarantees for the number of samples N for which this approach leads to a successful cubature construction with high probability, for a class of $\boldsymbol{\varphi}$ of practical interest.

Hypercontractivity. Our main tool is hypercontractivity. This allows us to prove the existence of a constant C_m' satisfying (mainly for $p = 4$)

$$\mathbb{E}[|f(X)|^p] \leq C_m' \mathbb{E}\left[|f(X)|^2\right]^{p/2}$$

uniformly for a large class of functions f, where X follows the product measure $\mu^{\otimes d}$. While hypercontractivity is classically studied for Gaussian, discrete, and uniform probability measures on hypercubes or hyperspheres [20, 125, 13, 14], we generalize it to function classes that have a certain graded structure.

Contribution. Our main result is to provide an upper bound for the number of samples N such that an N-point i.i.d. sample of random vectors contains the expectation in its convex hull, i.e. the event (3.2) occurs, with a reasonable probability. Although the connection between the bound for N and the hypercontractivity of the given random vector/function class has implicitly been proven in Chapter 2 in the form of Theorem 2.17, generic conditions for having a good hypercontractivity constant and why the magnitude of required N becomes reasonably small in practice have not been established or understood.

In this chapter, we address these questions by

- extending the hypercontractivity for the Wiener chaos to what we call generalized random polynomials (Section 3.3) and
- showing that this extension naturally applies to important examples in numerical analysis including classical cubature, cubature on Wiener space, and kernel quadrature (Section 3.4).

We explain the intuition behind these points by introducing Theorem 3.1 and Example 3.2:

Theorem 3.1 (informal). *Let μ be a probability measure on $\mathcal{X}$. Suppose we have a "natural" function class*

$$\mathcal{F} = \bigoplus_{d\ge1} \bigcup_{m\ge0} \mathcal{F}_{d,m},$$

where $\mathcal{F}_{d,m}$ denotes a set of functions from $\mathcal{X}^d$ to $\mathbb{R}$ of "degree" up to m. Then, under some integrability assumptions, there exists for every m a constant $C_m = C_m(\mu, \mathcal{F}) > 0$ such that the following holds:

> *Let d and D be two positive integers and $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_D) : \mathcal{X}^d \to \mathbb{R}^D$ with $\varphi_1, \dots, \varphi_D \in \mathcal{F}_{d,m}$. Then, for all integers $N \ge C_m D$, we have*
>
> $$\mathbb{P}(\mathbb{E}[\boldsymbol{\varphi}(X)] \in \operatorname{conv}\{\boldsymbol{\varphi}(X_1), \dots, \boldsymbol{\varphi}(X_N)\}) \ge \frac{1}{2},$$
>
> *where $X, X_1, \dots, X_N$ are i.i.d. samples from the product measure $\mu^{\otimes d}$ on $\mathcal{X}^d$.*

Example 3.2. Although the above statement is somewhat abstract, the assumption of a "natural" function class covers the following important examples:

- **Classical Cubature** [161]: μ is a probability measure with finite m moments and $\mathcal{F}_{d,m}$ is the space of d-variate polynomials up to degree m .
- **Cubature on Wiener space** [110]: μ is the Wiener measure and $\mathcal{F}_{d,m}$ is spanned by up to m-times iterated Ito–Stratonovich integrals.
- **Kernel quadrature** [84] (also see the next chapter): μ is a probability measure on set $\mathcal{X}$ that carries a positive definite kernel k and $\mathcal{F}_{d,m}$ is spanned by some test functions suitable to $k^{\otimes d}$, e.g., eigenfunctions down to some eigenvalue of the integral operator $g \mapsto \int k^{\otimes d}(\cdot, x)g(x)\,\mathrm{d}\mu^{\otimes d}(x)$, where $k^{\otimes d}$ is a tensor product kernel.

Related work. If the measure μ has finite support, the problem (3.1) is also known as recombination. While in this case, the existence follows immediately from Caratheodory's theorem, the design of efficient algorithms to compute the cubature measure is more recent; we mention efficient deterministic algorithms [103, 163, 111] and randomized speedups [34]. For non-discrete measures, the majority of the cubature constructions are typically limited to algebraic approaches that cannot apply to general situations. Related to our convex hull approach but different, is a line of research aiming at constructing general cubature formulas with positive weights by using least-squares instead of the random convex hull approach [55, 116]. As their theory is on the positivity of the resulting cubature formula given by solving a certain least squares problem, requires more (or efficiently selected) points than that needed for simply obtaining a positively weighted cubature.

Hypercontractivity is the key technical tool for our estimates. Although it is a classic tool in probability, its use seems to be novel in the context of cubature resp. random convex hull problems. Somewhat related to the special case of kernel quadrature, [115] proves a generalization error bound for kernel ridge regression with random features, however, hypercontractivity appears there just as a technical assumption. Further, for low-degree polynomials of a sequence of random variables, Kim and Vu [90], Schudy and Sviridenko [149] give similar estimates on their higher order moments, but they mainly estimate the concentration of the moments, and do not generally analyze the kurtosis-type values appearing in the hypercontractivity.

Outline. In Section 3.2, we briefly explain recent results on random convex hulls, and give some assertions that additionally follow from them. In Section 3.3, we introduce the Gaussian hypercontractivity and show its generalization that is suitable for multivariate cubatures. Section 3.4 gives some applications of Gaussian/generalized hypercontractivity to random convex hulls with product structure, including cubature on Wiener space and kernel quadrature. The chapter is concluded in Section 3.5. All the omitted proofs are given in Appendix 3.A.

3.2 Random convex hulls

Our main interest is the probability of the event (3.2) that the mean is contained in the random convex hull. To quantify this probability, it turns out to be more convenient to study a more general problem. Therefore we define

Definition 3.3. *Let X be a D-dimensional random vector and $X_1, X_2, \dots$ be iid copies of X. For every integer $N > 0$ and $\theta \in \mathbb{R}^D$ define*

$$p_{N,X}(\theta) := \mathbb{P}(\theta \in \operatorname{conv}\{X_1, \dots, X_N\}) \ \text{ and } N_X(\theta) := \inf\{N \mid p_{N,X}(\theta) \geq 1/2\}.$$

Both of these quantities are classically studied for symmetric X by Wendel [178], but more recently sharp inequalities for general X [172] (also Chapter 2 of the thesis) as well and calculations on the Gaussian case [81] have been established. Note that, by considering m independent copies of $(X_1, \dots, X_N)$, we have $1 - p_{mN,X}(\theta) \geq (1 - p_{N,X}(\theta))^m$, and so knowing $N_X(\theta)$ also yields some high-probability bounds.

Our main interest is the case $\theta = \mathbb{E}[X]$. We can bound $N_X(\mathbb{E}[X])$ if the distribution of X satisfies some good properties including symmetry and log-concavity:

Proposition 3.4. *For a D-dimensional random vector X for which $\mathbb{E}[X]$ exists, we have:*

(a) If the distribution of X is symmetric about $\mathbb{E}[X]$, then $N_X(\mathbb{E}[X]) \leq 2D$.

(b) If the distribution of X is log-concave, then $N_X(\mathbb{E}[X]) \leq \lceil 3eD \rceil$.

Here, (a) is a well-known result [178] while (b) follows from a combination of Caplin and Nalebuff [24] and Theorem 2.13; see Section 3.A.1 for details. However, what we use for our main results is the following reformulation of Theorem 2.17. We give its proof in Appendix 3.A.2 for completeness.

Corollary 3.5. *Let X be any D-dimensional random vector with $\mathbb{E}[\|X\|^3] < \infty$. If a constant $K > 0$ satisfies $\|c^\top(X - \mathbb{E}[X])\|_{L^3} \leq K\|c^\top(X - \mathbb{E}[X])\|_{L^2}$ for all $c \in \mathbb{R}^D$, then we have*

$$N_X(\mathbb{E}[X]) \leq 17(1 + 9K^6/4)D.$$

This result recovers a sharp bound $N_X(\mathbb{E}[X]) = \mathcal{O}(D)$ up to constant for a Gaussian, where we have detailed information about the marginals. The sort of inequality assumed in the statement is also called Khintchin's inequality (see e.g., [94, 47]) and there are known values of K for a certain class of X such as log-concave distributions [108, Theorem 5.22] or a Rademacher vector. Indeed, we can easily show the following estimate under a clear independence structure:

Proposition 3.6. *Let $X = (X_1, \ldots, X_D)^\top$ be a D-dimensional random vector whose coordinates are independent and identically distributed. If $\mathbb{E}[X_1] = 0$ and $\|X_1\|_{L^4} \leq K\|X_1\|_{L^2}$ holds for a constant $K > 0$, then we have $\|c^\top X\|_{L^4} \leq K\|c^\top X\|_{L^2}$ for all $c \in \mathbb{R}^D$.*

While such an explicit independence yields $N_X(\mathbb{E}[X]) = \mathcal{O}(D)$, we can see that we can go much further by carefully looking at how one can prove hypercontractivity in Gaussian Wiener chaos. In the following section, we generalize the whole argument and provide natural conditions for X to achieve $N_X(\mathbb{E}[X]) = \mathcal{O}(D)$.

3.3 Hypercontractivity

The previous section provides bounds on N_X but the assumptions–log-concavity or coordinate-wise independence–are too strong for many applications. We now develop an approach via hypercontractivity; this results in bounds that apply under much less strict assumptions.

Hypercontractivity: the Gaussian case. It is instructive to briefly review the classical results for Gaussian measures by following Janson [79] since we need several generalizations of this.

Theorem 3.7 (Wiener Chaos Decomposition). *Let H be a Gaussian Hilbert space*[1] *on a probability space $(\Omega, \mathcal{G}, \mathbb{P})$ and let $\sigma(H)$ be the σ-algebra generated by H. Then*

$$L^2(\Omega, \sigma(H), \mathbb{P}) = \bigoplus_{n=0}^{\infty} H^{(n)},$$

where $H^{(n)} := \overline{P_n(H)} \cap P_{n-1}(H)^{\perp}$ with

$$P_n(H) \coloneqq \{f(Y_1, \dots, Y_m) \mid f \text{ is a polynomial of degree} \le n,\ Y_1, \dots, Y_m \in H,\ m \ge 1\}$$

with $P_{-1}(H) \coloneqq \{0\}$ and $\overline{P_n(H)}$ denotes the completion in $L^2(\Omega, \mathcal{G}, \mathbb{P})$.

Hence, for each $X \in L^2(\Omega, \sigma(H), \mathbb{P})$, we have a unique decomposition $X = \sum_{n=0}^{\infty} X_n$ such that $X_n \in H^{(n)}$. The simplest case is expanding a random variable Y (measurable by a Gaussian variable X) with Hermite polynomials of X.

Theorem 3.8 (Hypercontractivity, [79], Theorem 5.8). *For $r \in [0, 1]$ denote*

$$T_r : L^2(\Omega, \sigma(H), \mathbb{P}) \to L^2(\Omega, \sigma(H), \mathbb{P}), \quad X \mapsto \sum_{n=0}^{\infty} r^n X_n.$$

If $p > 2$ and $0 < r \le (p-1)^{-1/2}$, then we have

$$\|T_r(X)\|_{L^p} \le \|X\|_{L^2}.$$

From this, we have the following moment bound on $\overline{P_n(H)}$, which is also referred to as hypercontractivity, see for example [127].

Theorem 3.9. *Let $n \ge 0$ be an integer. For each $p > 2$, we have*

$$\|X\|_{L^p} \le (p-1)^{n/2} \|X\|_{L^2}, \qquad X \in \overline{P_n(H)}.$$

[1] A Gaussian Hilbert space is a closed linear subspace of $L^2(\Omega, \mathcal{G}, \mathbb{P})$ whose elements all follow Gaussian distributions, where $\mathcal{G}$ is the σ-algebra of the given probability space.

Proof. Let $X = \sum_{m=0}^{n} X_m$ with $X_m \in H^{(m)}$. For $0 < r \leq (p-1)^{-1/2}$, by Theorem 3.8, we have

$$\|X\|_{L^p}^2 = \left\| T_r \left(\sum_{m=0}^{n} r^{-m} X_m \right) \right\|_{L^p}^2 \leq \left\| \sum_{m=0}^{n} r^{-m} X_m \right\|_{L^2}^2 = \sum_{m=0}^{n} r^{-2m} \|X_m\|_{L^2}^2 \leq r^{-2n} \|X\|_{L^2}^2.$$

We obtain the conclusion by letting $r = (p-1)^{-1/2}$. □

We included the proof since we are going to generalize it in the following.

Hypercontractivity for generalized random polynomials. The phenomenon of hypercontractivity is not limited to the Gaussian setting. Indeed, the hypercontractivity of operators on the space of boolean functions (i.e., $\{-1,1\}^n \to \mathbb{R}$) associated with the uniform measure was established even before the Gaussian case [20, 154]. Our focus is to obtain estimates analogous to Theorem 3.9 when a graded class of test function is given; we refer to such a class as generalized random polynomials.

Definition 3.10. *Under a probability space $(\Omega, \mathcal{G}, \mathbb{P})$, a triplet $G = (Y, Q, \lambda)$ is called GRP if it satisfies the following conditions:*

- *Y is a random variable taking values in a topological space $\mathcal{X}$.*
- *$Q = (Q_m)_{m=0}^{\infty}$ is a nondecreasing sequence of linear spaces of $L^2(\mathbb{P}_Y)$-integrable functions $\mathcal{X} \to \mathbb{R}$. Namely, if we let $Q_m(Y) := \{f(Y) \mid f \in Q_m\}$, then each Q_m is a linear subspace of $L^2(\mathbb{P})$, with $Q_0 \subset Q_1 \subset \cdots \subset L^2(\mathbb{P})$. We additionally assume Q_0 is the set of constant functions.*
- *$\lambda = (\lambda_m)_{m=0}^{\infty}$ satisfies $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$.*

If G is a GRP, we also define $\widetilde{\deg}_G X := \inf\{1/\lambda_m \mid m \geq 0,\, X \in \overline{Q_m(Y)}\}$.

Intuitively, each Q_m is a generalization of degree-m polynomials and $\widetilde{\deg}_G$ indicates the "degree" of such functions (though Y plays a role in the latter). In the setting of actual polynomials like Wiener chaos, we can define $\lambda_m = b^{-m}$ for a certain $b > 1$, and then we have $\deg X = \log_b \widetilde{\deg}_G X$ for the usual degree of X as a random polynomial.

Definition 3.11. *Let $G = (Y, Q, \lambda)$ be a GRP. We define*

$$H_m(Y) \coloneqq \overline{Q_m(Y)} \cap Q_{m-1}(Y)^\perp$$

in terms of $L^2(\mathbb{P})$ where $Q_{-1}(Y) \coloneqq \{0\}$ and

$$H_\infty \coloneqq L^2(\Omega, \sigma(Y), \mathbb{P}) \cap \big(\bigcup_{m=0}^{\infty} Q_m(Y)\big)^\perp.$$

We refer

$$L^2(\Omega, \sigma(Y), \mathbb{P}) = \left(\bigoplus_{m=0}^{\infty} H_m(Y)\right) \oplus H_\infty(Y)$$

as the orthogonal decomposition associated with G.

Definition 3.12. *Let $G = (Y, Q, \lambda)$ be a GRP. The operator $T(G)$ is defined as*

$$T(G) : L^2(\Omega, \sigma(Y), \mathbb{P}) \to L^2(\Omega, \sigma(Y), \mathbb{P}), \quad X \mapsto \sum_{m=0}^{\infty} \lambda_m X_m,$$

where $(X_m)_{m \in \mathbb{N} \cup \infty}$ with $X_m \in H_m(Y)$ is the orthogonal decomposition of X associated with the GRP G. By letting $T(G)^s X = \sum_{m=0}^{\infty} \lambda_m^s X_m$ for $s > 0$, we say that a GRP $G = (Y, Q, \lambda)$ is $(2, p; s)$-hypercontractive if

$$\|T(G)^s X\|_{L^p} \leq \|X\|_{L^2}, \qquad X \in L^2(\Omega, \sigma(Y), \mathbb{P}).$$

Thus, if G is $(2, p; s)$-hypercontractive, it is $(2, p; t)$-hypercontractive for all $t \geq s$ as $T(G)^{t-s}$ is a contraction in L^2. The formulation of G associated with "degree" concept given by λ then naturally extends to the multivariate case.

Definition 3.13. *We call a set of d GRPs, $G^{(i)} = (Y^{(i)}, Q^{(i)}, \lambda^{(i)})$ for $i = 1, \ldots, d$ independent, if the $Y^{(i)}$'s are independent random variables taking values in $\mathcal{X}^{(i)}$'s. For d independent GRPs, their product is a GRP $G = (Y, Q, \lambda)$ that is defined as*

- $Y = (Y^{(1)}, \ldots, Y^{(d)}) \in \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(d)}$.
- λ_m *is the $(m+1)$-th largest value in the set* $\left\{\prod_{i=1}^d \lambda_{m_i}^{(i)} \,\middle|\, \lambda_{m_i}^{(i)} \in \lambda^{(i)},\ i = 1, \ldots, d\right\}$.
- $Q_m = \operatorname{span}\left\{f : (x_1, \ldots, x_d) \mapsto \prod_{i=1}^d f_i(x_i) \,\middle|\, f_i \in Q_{m_i}^{(i)},\ \prod_{i=1}^d \lambda_{m_i}^{(i)} \leq \lambda_m\right\}$.

As $Q_m(Y) \subset L^2$ it follows from independence for each m that $G = (Y, Q, \lambda)$ is indeed a GRP. We also denote it by $G = G^{(1)} \otimes \cdots \otimes G^{(d)}$.

Example 3.14. *Consider the case when $Q_m^{(i)}$ are degree-m polynomials of $Y^{(i)}$ and $\lambda_m^{(i)} = t^m$ for some $t \in (0,1)$ independent of i. This shows that the product GRP generalizes the multivariate random polynomials. Also, when $Y^{(i)}$ are i.i.d. and $(Q^{(i)}, \lambda^{(i)})$ are the same for all $i = 1, \ldots, d$, then we say $G^{(i)}$ are i.i.d. and we can particular write $G \simeq (G^{(1)})^{\otimes d}$.*

A straightforward generalization follows from the classical way of proving hypercontractivity. This turns out to be useful for treating multivariate hypercontractivity of our GRP setting.

Theorem 3.15. *Let $r \in (0,1]$ and $p > 2$. If d independent GRPs $G^{(1)}, \ldots, G^{(d)}$ are all $(2,p;s)$-hypercontractive, then their product $G = G^{(1)} \otimes \cdots \otimes G^{(d)}$ is also $(2,p;s)$-hypercontracitve.*

Remark 3.1. We only use the $(2,p;s)$-hypercontractivity in this chapter, but we can also deduce the same results for the general $(q,p;s)$-hypercontractivity with $1 \le q \le p < \infty$ (for the operator norm of $L^q \to L^p$), analogous to e.g. Janson [79].

The following is analogous to Theorem 3.9 and the proof is almost identical.

Proposition 3.16. *Let $s > 0$ and $p > 2$. If G is a GRP that is $(2,p;s)$-hypercontractive, then we have $\|X\|_{L^p} \le (\widetilde{\deg_G X})^s \|X\|_{L^2}$ for all $X \in L^2$.*

Remark 3.2. Although we have treated general GRPs $G = (Y, Q, \lambda)$ in these propositions, we are basically only interested in the moment inequality for X up to some "degree" fixed beforehand (in the case of Wiener chaos, it suffices to treat $P_n(H)$ for some finite n to obtain Theorem 3.9). Thus, our main interest is in "finite" GRPs, satisfying $Q_n = Q_{n+1} = \cdots$ for some n, and their product in practice, which might be better for readers to have in mind when reading the next proposition.

We next show the following "converse" result for the relation of the hypercontractivity and moment estimate for a (truncated) GRP when $p = 4$.

Proposition 3.17. *Let $G = (Y, Q, \lambda)$ be a GRP. Suppose there exists a $s > 0$ such that*

$$\|X_m\|_{L^4} \leq \lambda_m^{-s} \|X_m\|_{L^2}, \qquad X_m \in H_m(Y)$$

holds for all m. If $t > s$ satisfies $\sum_{m \geq 1} \lambda_m^{t-s} \leq 1/\sqrt{3}$ and $\lambda_1^t \leq 1/2$, then G is $(2, 4; t)$-hypercontractive.

By using this, we can also prove the following as a non-quantitative result.

Theorem 3.18. *Let $K > 0$ and G be a GRP such that the space $\{X \in L^2 \mid \widetilde{\deg}_G X \leq K\}$ is included in $L^4(\Omega, \mathcal{G}, \mathbb{P})$ and finite-dimensional. Then, there exists a constant $C = C(G, K)$ such that for an arbitrary d, $\|X\|_{L^4} \leq C\|X\|_{L^2}$ holds if we have $\widetilde{\deg}_{G^{\otimes d}} X \leq K$.*

3.4 Applications

The generality of Proposition 3.17 and Theorem 3.18 allows quantifying the number of samples resp. probability of success of the random convex hull approach to the problem of cubature. Concretely, we give formal statements of Theorem 3.1 for various cubature constructions: (i) Classical Cubature, (ii) Cubature on Wiener Space, (iii) Kernel Quadrature.

3.4.1 Classical polynomial cubatures

When the GRP G are actual random polynomials, we recover the following result.

Corollary 3.19. *Let m be a positive integer and $X^{(1)}, X^{(2)}, \ldots$ be i.i.d. real-valued random variables with $\mathbb{E}\big[|X^{(1)}|^{4m}\big] < \infty$. Then, there exists a constant $C_m > 0$ such that*

$$\|f(X^{(1)}, \ldots, X^{(d)})\|_{L^4} \leq C_m \|f(X^{(1)}, \ldots, X^{(d)})\|_{L^2}$$

for any positive integer d and any polynomial $f : \mathbb{R}^d \to \mathbb{R}$ with degree up to m.

Proof. By introducing a truncated GRP given by a random variable $X^{(1)}$, function spaces Q_i of univariate polynomials up to degree i, and $\lambda_i = 2^{-i}$ for $0 \leq i \leq m$, we can apply Theorem 3.18 to obtain the desired result. □

If we combine this with Corollary 3.5, we obtain the following result for polynomial cubatures:

Corollary 3.20. *Let $m \geq 1$ be an integer and $X^{(1)}, X^{(2)}, \dots$ be i.i.d. real-valued random variables with $\mathbb{E}\big[|X^{(1)}|^{4m}\big] < \infty$. Then, there exists a constant $C_m > 0$, depending on the law of $X^{(1)}$, such that the following holds:*

> *Let $d \geq 1$ be an integer and $\boldsymbol{\varphi} : \mathbb{R}^d \to \mathbb{R}^D$ be a D-dimensional vector-valued function such that each coordinate is given by a polynomial up to degree m. If we let $\boldsymbol{X}_1^{(1:d)}, \boldsymbol{X}_2^{(1:d)}, \dots$ be independent copies of $\boldsymbol{X}^{(1:d)} = (X^{(1)}, \dots, X^{(d)})$, we have*
>
> $$\mathbb{P}\Big(\mathbb{E}\Big[\boldsymbol{\varphi}(\boldsymbol{X}^{(1:d)})\Big] \in \operatorname{conv}\{\boldsymbol{\varphi}(\boldsymbol{X}_1^{(1:d)}), \dots, \boldsymbol{\varphi}(\boldsymbol{X}_N^{(1:d)})\}\Big) \geq \frac{1}{2}$$
>
> *for all integers $N \geq C_m D$.*

3.4.2 Cubature on Wiener space

Cubature on Wiener space [110] is a weak approximation scheme for stochastic differential equations; at the heart of it is constructing a finite measure on the space of paths, such that the expectation of their first m-times iterated integrals matches those of Brownian motion. Some algebraic constructions are known for degree $m = 3, 5$ [110] as well as $m = 7$ [126]. The random convex hull approach applies in principle for any m, however, a caveat is that the discretization of paths becomes an issue in particular for high values of m; some experimental results are available in [68] for constructing them by using random samples of piecewise linear approximations of Brownian motion. In this section, we use hypercontractivity to estimate the number of samples needed in this approach to cubature via sampling.

Random convex hulls of iterated integrals. For a bounded-variation (BV) path $x = (x^0, \dots, x^d) : [0,1] \to \mathbb{R}^{d+1}$ and a d-dimensional standard Brownian motion $B = (B^1, \dots, B^d)$ with $B_t^0 := t$, we define the iterated integrals as

$$I^\alpha(x) := \int_{0<t_1<\cdots<t_k<1} \mathrm{d}x_{t_1}^{\alpha_1} \cdots \mathrm{d}x_{t_k}^{\alpha_k}, \qquad I^\alpha(B) := \int_{0<t_1<\cdots<t_k<1} \circ\, \mathrm{d}B_{t_1}^{\alpha_1} \cdots \circ\, \mathrm{d}B_{t_k}^{\alpha_k},$$

where the latter is given by the Stratonovich stochastic integral. Then, a degree m cubature formula on Wiener space for d-dimensional Brownian motion is a set of BV paths $x_1, \dots, x_n : [0,1] \to \mathbb{R}^{d+1}$ and convex weights $w_1, \dots, w_n$ such that $\sum_{i=1}^n w_i I^\alpha(x_i) = \mathbb{E}[I^\alpha(B)]$ for all multi-indices $\alpha = (\alpha_1, \dots, \alpha_k) \in \bigcup_{\ell \geq 1}\{0, 1, \dots, d\}^\ell$ with $\|\alpha\| := k + |\{j \mid \alpha_j = 0\}| \leq m$.

Indeed, if we consider the Gaussian Hilbert space given by

$$H := \left\{ \sum_{i=1}^d \int_0^1 f_i(t) \, \mathrm{d}B_t^i \;\middle|\; f_1, \dots, f_d \in L^2([0,1]) \right\},$$

the iterated integral $I^\alpha(B)$ with $\|\alpha\| \leq m$ is in the m-th Wiener chaos $\overline{P_m(H)}$ (see Section 3.3) as it can be expressed as a limit of polynomials of increments of B. We thus have the hypercontractivity given in Theorem 3.9 and the following:

Corollary 3.21. *Let $d, m \geq 1$ be integers and B be a d-dimensional Brownian motion. Then, for an arbitrary linear combination $X = \sum_{\|\alpha\| \leq m} c_\alpha I^\alpha(B)$ with $c_\alpha \in \mathbb{R}$, we have $\|X\|_{L^3} \leq 2^{m/2} \|X\|_{L^2}$.*

As the bound is independent of the dimension d of the underlying Brownian motion, we have the following version of Theorem 3.1 by combining it with Corollary 3.5 as follows:

Corollary 3.22. *Let $d, m \geq 1$ be integers and $B, B_1, B_2, \dots$ be independent standard d-dimensional Brownian motions. Then, if $\boldsymbol{\varphi}(B)$ is a D-dimensional random vector such that each coordinate is given by a linear combination of $(I^\alpha(B))_{\|\alpha\| \leq m}$, then we have*

$$\mathbb{P}(\mathbb{E}[\boldsymbol{\varphi}(B)] \in \operatorname{conv}\{\boldsymbol{\varphi}(B_1), \dots, \boldsymbol{\varphi}(B_N)\}) \geq \frac{1}{2}$$

for all integers $N \geq 17(1 + 18 \cdot 8^{m-1})D$.

The above allows to choose the number of candidate paths that need to be sampled. However, as mentioned above, one challenge that is specific to cubature on the space of paths is that one cannot sample a true Brownian trajectory which leads to an additional discretization error. However, we conjecture that the number of random samples divided by D and the number of time partitions for piecewise linear approximation in constructing cubature on Wiener space can be independent of the underlying dimension d.

Remark 3.3. One can also apply these estimates to fractional Brownian motion [134], though we also need to obtain the exact expectations of iterated integrals of fractional Brownian motion (we can find some results on the Ito-type iterated integrals without the time integral by $B_t^0 = t$ in the literature [11, Theorem 31]).

3.4.3 Kernel quadrature for product measures

Recall the kernel quadrature problem introduced in Section 1.2; we are given a positive definite kernel k and a Borel probability measure μ on a space $\mathcal{X}$. Our aim is to find a good kernel quadrature rule, a set of points $x_i \in \mathcal{X}$ and weights $w_i \in \mathbb{R}$ such that $Q_n = \sum_{i=1}^n w_i \delta_{x_i}$ with the small worst-case error $\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)$, which we might just denote by $\mathrm{wce}(Q_n)$. We call a kernel quadrature convex if Q_n is a probability measure, i.e., $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$.

In this section, we see that the complexity of randomized kernel quadrature algorithm (a version of Algorithm 4.2 in the next chapter) is related to the hypercontractivity discussed in this chapter, and give some bounds based on GRPs when RKHS has a product structure.

Tensor product kernels. With d space-kernel pairs $(\mathcal{X}_1, k_1), \dots, (\mathcal{X}_d, k_d)$, the *tensor product kernel* on the product space $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ is defined as

$$(k_1 \otimes \cdots \otimes k_d)(x, y) := \prod_{i=1}^d k_i(x_i, y_i),$$

where $x = (x_1, \dots, x_d), y = (y_1, \dots, y_d) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. This is indeed the reproducing kernel of the tensor product $\mathcal{H}_{k_1} \otimes \cdots \otimes \mathcal{H}_{k_d}$ in terms of RKHS [18]. The most important example of this construction is when the underlying d kernels are the same, $k^{\otimes d} = k \otimes \cdots \otimes k$. Given a probability measure μ in the (conceptually univariate) space $\mathcal{X}$, constructing a kernel quadrature for $\mu^{\otimes d}$ with respect to $k^{\otimes d}$ is a natural multivariate extension of kernel quadrature that is widely studied in the literature [131, 82, 6, 84], and corresponds to high-dimensional QMCs [41]. While we will ultimately consider kernel quadrature for $(k^{\otimes d}, \mu^{\otimes d})$, let us start from the "univariate" (k, μ) in the following.

Mercer-like expansions and quadrature. Let us consider the Mercer-type expansion:

$$k(x, y) = \sum_{\ell=1}^{\infty} \sigma_\ell e_\ell(x) e_\ell(y), \tag{3.3}$$

where we suppose $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ and $e_\ell \in L^2(\mu)$ (not necessarily normalized). Although $(\sigma_\ell, e_\ell)_{\ell=1}^\infty$ are given by normalized eigenpairs of the integral operator $\mathcal{K} : f \mapsto \int_{\mathcal{X}} k(\cdot, y) f(y) \,\mathrm{d}\mu(y)$ in the case of Mercer expansion [160], we can also use other expansions such as power-series expansion [185] and the Nyström-based (truncated) expansion (discussed in the next two chapters). The following proposition, which will formally be proven in a more general form as Proposition 4.8 in the next chapter, shows how a finite-dimensional cubature in the sense of (3.1) yields some meaningful kernel quadrature.

Proposition 3.23. *Let $Q_n = (w_i, x_i)_{i=1}^n$ be a convex kernel quadrature satisfying $\int_{\mathcal{X}} e_\ell(x) \,\mathrm{d}\mu(x) = \sum_{i=1}^n w_i e_\ell(x_i)$ for each $\ell = 1, \ldots, n-1$. Then, by letting $r_n(x) := \sum_{m=n}^\infty \sigma_m e_m(x)^2$, we have $\mathrm{wce}(Q_n)^2 \leq 4 \sup_{x \in \mathcal{X}} r_n(x)$.*

We can have more favorable bounds on $\mathrm{wce}(Q_n)$ by assuming more (see the next chapter), but the important fact here is that the event (3.2) for a vector-valued $\boldsymbol{\varphi}$ given by "basis" functions $e_1, \ldots, e_{n-1}$ enables us to construct an interesting numerical scheme. A similar approach, specialized to a Gaussian kernel over a Gaussian measure can be found in [84]. Given a Mercer-like expansion (3.3), we can also consider the multivariate version

$$k^{\otimes d}(x, y) = \sum_{\ell_1, \ldots, \ell_d = 1}^{\infty} \sigma_{\ell_1} \cdots \sigma_{\ell_d} (e_{\ell_1} \otimes \cdots \otimes e_{\ell_d})(x)(e_{\ell_1} \otimes \cdots \otimes e_{\ell_d})(y), \tag{3.4}$$

and the same result as Proposition 3.23 holds for the properly reordered expansion. For the interaction between the convergence rate and the dimension d in the case of Mercer expansion, Bach [6, Section 3.4] provides some informative examples.

As the construction of such Q_n in Proposition 3.23 for general k and μ relies on random sampling, we want to estimate $N_{\boldsymbol{\varphi}(X)}(\mathbb{E}[\boldsymbol{\varphi}(X)])$ for $X \sim \mu$ and $\boldsymbol{\varphi} = (e_1, \ldots, e_{n-1})$, where our main interest lies in the multivariate case despite using the univariate notation for simplicity.

From RKHS to GRP. To make it compatible with the framework of GRPs introduced in the previous section, we further assume the following condition, which ensures that the kernel is in an appropriate scaling.

Assumption A. *The expansion* (3.3) *converges pointwise,* $\sum_{\ell=1}^\infty \sigma_\ell < \infty$*,* $\sigma_1 \le 1$*, and the strict inequality* $\sigma_\ell < 1$ *holds if* $e_\ell \in L^2(\mu)$ *is not constant.*

Under Assumption A, we can naturally define a GRP $G = (Y, Q, \lambda)$ with Y following μ, $Q_m = \operatorname{span}\{1, e_1, \dots, e_m\}$ and $\lambda_m = \sigma_m$ for $m \ge 1$. Note that it violates the condition $\lambda_1 < 1$ if $\sigma_1 = 1$ and e_1 is constant, but in that case we can simply decrement all the indices of (Q_m, λ_m) by one. We call it the *natural GRP* for k (with the expansion) and μ, and we denote it by $G = G_{k,\mu}$.

Remark 3.4. The scaling given in Assumption A is essential to the hypercontractivity under the framework of tensor product kernels when considering "eigenspace down to some eigenvalue". Indeed, if $\sigma_\ell \ge 1$ for some nonconstant e_ℓ, we have

$$\frac{\|e_\ell^{\otimes d}\|_{L^p(\mu^{\otimes d})}}{\|e_\ell^{\otimes d}\|_{L^2(\mu^{\otimes d})}} = \left(\frac{\|e_\ell\|_{L^p(\mu)}}{\|e_\ell\|_{L^2(\mu)}}\right)^d$$

for $p > 2$, which increases exponentially as d grows, whereas the corresponding $(\sigma_\ell)^d$ is lower bounded by 1. So the hypercontractivity in our sense never gets satisfied if $\sigma_\ell \ge 1$ for a nonconstant e_ℓ.

By introducing GRPs as above, we can prove the following statement, written without GRPs.

Proposition 3.24. *Let* k *satisfy Assumption A and* $Y_1, Y_2, \dots$ *independently follow* μ*. For each* $\varepsilon > 0$*, define a set of random variables as*

$$S(\varepsilon) := \operatorname{span}(\{1\} \cup \{e_{\ell_1}(Y_{m_1}) \cdots e_{\ell_d}(Y_{m_d}) \mid d \ge 1,\, m_1 < \cdots < m_d,\, \sigma_{\ell_1} \cdots \sigma_{\ell_d} \ge \varepsilon\}).$$

Then, if $\|e_\ell(Y_1)\|_{L^4} < \infty$ *holds for all* ℓ *with* $\sigma_\ell \ge \varepsilon$*, then there is a constant* $C_\varepsilon > 0$ *such that* $\|X\|_{L^4} \le C_\varepsilon \|X\|_{L^2}$ *for all* $X \in S(\varepsilon)$*.*

Proof. The finiteness of the dimension of "eigenspace" for Y_1, i.e, the finiteness of the number of ℓ satisfying $\sigma_\ell \ge \varepsilon$ follows from $\sum_{\ell=1}^\infty \sigma_\ell < \infty$ in Assumption A. Thus, Theorem 3.18 gives the conclusion. □

If we only had $Y_1, \dots, Y_d$, then $S(\varepsilon)$ would correspond to the truncation of the d-variate expansion (3.4). So this assertion includes a hypercontractivity statement for an "eigenspace" of $k^{\otimes d}$ and $\mu^{\otimes d}$ given the expansion (3.4). However, we can go further to a quantitative statement by imposing another assumption in the case of actual Mercer expansion.

Quantitative bounds for Mercer expansion. We first set up two additional assumptions for obtaining a quantitative statement. We shall discuss The following assumption says that (3.3) is actually the Mercer expansion.

Assumption B. *$(e_\ell)_{\ell=1}^\infty$ and $(\sqrt{\sigma_\ell} e_\ell)_{\ell=1}^\infty$ are orthonormal sets in $L^2(\mu)$ and $\mathcal{H}_k$, respectively.*

Mild conditions already imply that Assumption B holds, e.g., $\operatorname{supp} \mu = \mathcal{X}$, k is continuous, and $x \mapsto k(x, x)$ is in $L^1(\mu)$ is sufficient, see [160]. Another assumption requires further orthogonality of these test functions against a constant function.

Assumption C. *The kernel k can be written as $k = 1 + k_0$, where $k_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel satisfying $\int_{\mathcal{X}} k_0(x, y) \, \mathrm{d}\mu(y) = 0$ for (μ-almost) all $x \in \mathcal{X}$.*

Under Assumption A, B, this is simply equivalent to e_1 being constant. This might seem artificial, but naturally arises in the following situations:

(a) $\mathcal{X}$ is a compact group and μ is its Haar measure. k is a positive definite kernel given as $k(x, y) = g(x^{-1}y)$, where $g : \mathcal{X} \to \mathbb{R}_{\geq 0}$ and $\int_{\mathcal{X}} g(x) \, \mathrm{d}\mu(x) = 1$.

(b) k_0 is a kernel called Stein kernel [129, 5] with appropriate scaling.

One theoretically sufficient condition for these assumptions can be described as follows:

Proposition 3.25. *Let $\mathcal{X}$ be compact metrizable and path-connected,* $\operatorname{supp} \mu = \mathcal{X}$, *and k be continuous and nonnegative. If $\int_{\mathcal{X}} k(x, y) \, \mathrm{d}\mu(y) = 1$ holds for all $x \in \mathcal{X}$, Assumption A–C hold.*

From this proposition, for instance, an appropriately scaled exponential/Gaussian kernel over the n-sphere with the uniform measure satisfies Assumption A–C.

Under these two assumptions, the operator $T(G_{k,\mu})$ in terms of GRPs corresponds to the integral operator $\mathcal{K} : f \mapsto \int_{\mathcal{X}} k(\cdot, y)f(y)\,\mathrm{d}\mu(y)$, so the situation becomes simpler. We can directly apply Proposition 3.17 by replacing λ's with σ's, but we also have the following sufficient conditions for the hypercontractivity without explicitly using the eigenvalue sequence. In the following, $\|\mathcal{K}_0\| := \sigma_2 < 1$ is the operator norm of $\mathcal{K}_0 : f \mapsto \int_{\mathcal{X}} k_0(\cdot, y)f(y)\,\mathrm{d}\mu(y)$ on $L^2(\mu)$, and $\mathrm{tr}(\mathcal{K}_0) := \int_{\mathcal{X}} k_0(x,x)\,\mathrm{d}\mu(x)$. We have the following quantitative condition for hypercontractivity.

Proposition 3.26. *Let $k = 1 + k_0$ satisfy Assumption A–C. When $\|\mathcal{K}_0\| > 0$, if $r, s \ge 1$ satisfy*

$$\|\mathcal{K}_0\|^{-(r+s)} \ge 2, \quad \|\mathcal{K}_0\|^{-(r-1)} \ge \sqrt{3}\,\mathrm{tr}(\mathcal{K}_0), \quad \|\mathcal{K}_0\|^{-(s-1)} \ge \|k_0\|_{L^4(\mu\otimes\mu)},$$

then $G_{k,\mu}$ is $(2,4;r+s)$-hypercontractive. In particular, if we have $\sup_{x\in\mathcal{X}}|k_0(x,x)| \le 1/\sqrt{3}$, then $G_{k,\mu}$ is $(2,4;2)$-hypercontractive.

Example 3.27 (Periodic Sobolev spaces over the torus.). *Following Bach [6], we consider periodic kernels over $[0,1]$. Therefore let $\mathcal{X} = [0,1]$, μ be the uniform distribution on $\mathcal{X}$, and define*

$$k_{r,\delta}(x,y) = 1 + \delta \cdot \frac{(-1)^{r-1}(2\pi)^{2r}}{(2r)!} B_{2r}(|x-y|) \tag{3.5}$$

for each positive integer s and $\delta \in (0,1)$, where B_{2r} is the $2r$-th Bernoulli polynomial [173]. $\delta = 1$ is assumed in the original definition, but it violates Assumption A (see also Remark 3.4). Albeit this slight modification, the kernel $k_{r,\delta}$ gives an equivalent norm to the periodic Sobolev space in the literature. For $\delta \in (0,1)$, $k_{r,\delta}$ satisfies Assumption A–C. The eigenvalues and eigenfunctions with respect to the uniform measure are known [6]; the eigenvalues are: 1 for the constant function, and δm^{-2r} for $c_m(\cdot) := \sqrt{2}\cos(2\pi m\,\cdot)$ and $s_m(\cdot) := \sqrt{2}\sin(2\pi m\,\cdot)$ for $m \ge 1, 2, \ldots$. We now apply Proposition 3.17 with (for sake of concreteness) $\delta = 1/3$. This gives $\|c_m\|_{L^4(\mu)} = \|s_m\|_{L^4(\mu)} = (3/2)^{1/4}$. Thus, to satisfy the condition of Proposition 3.17, it suffices for $s < t$ to satisfy $3^s \ge (3/2)^{1/4}$, $\delta^{t-s}\zeta(2r(t-s))$, $3^t \ge 2$, where

ζ is Riemann's zeta function. Hence a simple numerical sufficient condition for this is $s = 0.1$ and $t = 1.1$ for $r = 1$, and $t = \log_3 2 \le 0.631$ for $r \ge 2$, which can be derived by letting $2r(t - s) \ge 2$. To sum up, in the case $r \ge 2$, we only need $\mathcal{O}(\lambda^{-0.631} D)$ times of sampling for meeting (3.2) *with probability over a half, if $X \sim \mu^{\otimes d}$ and each coordinate of $\boldsymbol{\varphi} : \mathcal{X}^d \to \mathbb{R}^D$ is in the eigenspace of the eigenvalue λ.*

3.5 Concluding remarks

We investigated the number of samples needed for the expectation vector to be contained in their convex hull from the viewpoint of product/graded structure. We demonstrated that we empirically only need $\mathcal{O}(D)$ times of sampling for the D-dimensional random vector in practical examples can partially be explained by the hypercontractivity in the Gaussian case as well as the generalized situation including random polynomials and product kernels. There are also interesting questions for further research; for example, although in the asymptotic $d \to \infty$ we established that the required number of sampling divided by D is independent of d, the constants are larger than what purely empirical estimates given in [66] (and the next chapter, where $10D$ is sufficient in practice). Another direction, is the case of cubature of Wiener space, as one cannot actually sample from Brownian motion and discretization errors propagate to higher order m; a promising research direction could be to study "approximate sampling" or consider unbiased simulations [74] for the iterated integrals.

Appendix for Chapter 3

3.A Proofs

3.A.1 Proof of Proposition 3.4(b)

For a D-dimensional random vector X, recall the following Tukey depth defined in Chapter 2:

$$\alpha_X(\theta) \coloneqq \inf_{c \in \mathbb{R}^D \setminus \{0\}} \mathbb{P}\big(c^\top (X - \theta) \le 0\big) . \tag{3.6}$$

We have shown $N_X(\theta) \leq \lceil 3D/\alpha_X(\theta)\rceil$ in Theorem 2.13.

The above can be used to provide a novel bound on $N_X(\mathbb{E}[X])$ for a general class of distributions called *log-concave* as in the statement of Proposition 3.4. A function $f : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is called *log-concave* if it satisfies

$$f(tx + (1-t)y) \geq f(x)^t f(y)^{1-t}$$

for all $x, y \in \mathbb{R}^d$ and $t \in [0,1]$. A probability distribution with a log-concave density is also called log-concave, and this class includes the multivariate Gaussian/exponential/Wishart distributions, the uniform distribution over a convex domain, and many more univariate common distributions [4, 21]. For the log-concave random vectors, the following result is known:

Theorem 3.28 ([24])**.** *If X is a d-dimensional random vector with log-concave density, then we have $\alpha_X(\mathbb{E}[X]) \geq 1/e$.*

The case, when X is uniform over a convex set, is proven in Grünbaum [59], and Lovász and Vempala [108, Section 5] gives simpler proofs than the original result in Caplin and Nalebuff [24].

Simply combining Theorem 3.28 and $N_X(\mathbb{E}[X]) \leq \lceil 3D/\alpha_X(\mathbb{E}[X])\rceil$ in Theorem 2.13 yields the desired result.

3.A.2 Proof of Corollary 3.5

Proof. Let $Y := X - \mathbb{E}[X]$. First assume that $V := \mathbb{E}\left[YY^\top\right]$ is nonsingular, then we have $\|c^\top Y\|_{L^2} = \sqrt{c^\top V c} = \|V^{1/2}c\|_2$. Thus, we have

$$\sup_{c\in\mathbb{R}^D\setminus\{0\}} \frac{\|c^\top Y\|_{L^3}}{\|c^\top Y\|_{L^2}} = \sup_{c\in\mathbb{R}^D\setminus\{0\}} \frac{\|(V^{1/2}c)^\top V^{-1/2}Y\|_{L^3}}{\|V^{1/2}c\|_2} = \sup_{a\in\mathbb{R}^D,\,\|a\|_2=1} \|a^\top V^{-1/2}Y\|_{L^3}.$$

Therefore, the assertion follows by observing that the sixth power of the right-hand side appears in the bound of Theorem 2.17 (by using Y instead of X).

We next consider the case when V is singular. We prove by induction on D. $D = 1$ with a singular V implies that X is almost surely a constant and $N_X(\mathbb{E}[X]) = 1$. Let us assume $D > 1$. Since V is singular, there exists a vector $u \in \mathbb{R}^D \setminus \{0\}$ such that $u^\top V u = 0$. Therefore, we have

$$0 = u^\top V u = u^\top \mathbb{E}\left[YY^\top\right] u = \mathbb{E}\left[(u^\top Y)^2\right]$$

and so $u^\top Y = 0$ almost surely. Therefore, there exists an index i such that $Y^{(i)}$ (i-th entry of Y) is (almost surely) determined by a linear combination of the other entries. Say, $Y^{(i)} = v^\top Y'$ holds almost surely with $v \in \mathbb{R}^{D-1}$ and Y' being the $(D-1)$-dimensional random vector given by omitting the i-th entry of Y. Let $A_i \in \mathbb{R}^{(D-1)\times D}$ be the linear map of omitting the i-th entry of the given vector. For any $c \in \mathbb{R}^D$, we have

$$c^\top Y = (A_i c)^\top Y' + c^{(i)} Y^{(i)} = (A_i c + c^{(i)} v)^\top Y'$$

almost surely, so we have the same constant K in the assumption of the corollary for Y and Y'. Now it suffices to prove $\alpha_Y(0) = \alpha_{Y'}(0)$ from $D - 1 \leq D$ and the induction hypothesis.

Indeed, let $Y_1, \dots, Y_N$ be independent copies of Y and $Y_1', \dots, Y_N'$ be given by omitting their i-th entries. The latter sequence is independent copies of Y'. Then, for $c_1, \dots, c_N \geq 0$ with $c_1 + \cdots + c_N = 1$, $\sum_{j=1}^N c_j Y_j = 0$ clearly implies $\sum_{j=1}^N c_j Y_j' = 0$. The contrary also holds almost surely, because, for the i-th entries $Y_1^{(i)}, \dots, Y_N^{(i)}$, we almost surely have

$$\sum_{j=1}^N c_j Y_j^{(i)} = \sum_{j=1}^N c_j v^\top Y_j' = v^\top \sum_{j=1}^N c_j Y_j' = 0$$

if $\sum_{j=1}^N c_j Y_j' = 0$. Therefore, we have $\alpha_{Y'}(0) = \alpha_Y(0) = \alpha_X(\mathbb{E}[X])$ and obtain the desired estimate for when the dimension is D. □

3.A.3 Proof of Proposition 3.6

Proof. It suffices to consider the case $\|X_1\|_{L^4} < \infty$. If we write $c = (c_1, \dots, c_D)^\top$, then by using independence, we have

$$\begin{aligned}\|c^\top X\|_{L^4}^4 = \mathbb{E}\left[(c^\top X)^4\right] &= \sum_{i=1}^D c_i^4 \mathbb{E}\left[X_i^4\right] + \sum_{1\leq i<j\leq D} c_i^2 c_j^2 \mathbb{E}\left[X_i^2\right]\mathbb{E}\left[X_j^2\right] \\ &\leq K^4 \sum_{i=1}^D c_i^4 \mathbb{E}\left[X_i^2\right]^2 + \sum_{1\leq i<j\leq D} c_i^2 c_j^2 \mathbb{E}\left[X_i^2\right]\mathbb{E}\left[X_j^2\right] \\ &\leq K^4 \left(\sum_{i=1}^D c_i^2 \mathbb{E}\left[X_i^2\right]\right)^2 \leq K^4 \mathbb{E}\left[(c^\top X)^2\right]^2,\end{aligned}$$

as we clearly have $K \geq 1$ (or $X = 0$ almost surely). □

3.A.4 Proof of Theorem 3.15

Proof. We give the proof by generalizing the proof of Lemma 5.3 in Janson [79].

It suffices to prove the statement for $d = 2$, as the product of GRPs is associative. Let $G^{(i)} = (Y^{(i)}, Q^{(i)}, \lambda^{(i)})$ for $i = 1, 2$ be independent GRPs. Let $H_m^{(i)}(Y^{(i)}) := \overline{Q_m^{(i)}(Y^{(i)})} \cap Q_{m-1}^{(i)}(Y^{(i)})^{\perp}$ for $i = 1, 2$. If we denote the product by $G = G^{(1)} \otimes G^{(2)}$. Then, for a random variable $X = \sum_{\ell,m} X_{\ell,m}$ with $X_{\ell,m} \in H_\ell^{(1)} \otimes H_m^{(2)}$, the operator $T(G)$ acts as

$$T(G)X = \sum_{\ell,m} \lambda_\ell^{(1)} \lambda_m^{(2)} X_{\ell,m}.$$

If each $X_{\ell,m}$ can be written as a finite sum $X_{\ell,m} = \sum_k X_{k,\ell,m}^{(1)} X_{k,\ell,m}^{(2)}$ with $X_{k,\ell,m}^{(1)} \in H_\ell^{(1)}(Y^{(1)})$ and $X_{k,\ell,m}^{(2)} \in H_m^{(2)}(Y^{(2)})$, then by using Minkowski's integral inequality [65] and the $(2, p; s)$-hypercontractivity of $G^{(1)}$ and $G^{(2)}$, we have

$$\begin{aligned}
\|T(G)^s X\|_{L^p} &= \mathbb{E}_{Y^{(1)}}\left[\mathbb{E}_{Y^{(2)}}\left[\left|\sum_{\ell,m} (\lambda_\ell^{(1)} \lambda_m^{(2)})^s X_{\ell,m}\right|^p\right]\right]^{1/p} \\
&= \mathbb{E}_{Y^{(1)}}\left[\mathbb{E}_{Y^{(2)}}\left[\left|\sum_{k,\ell,m} (\lambda_\ell^{(1)})^s X_{k,\ell,m}^{(1)} (\lambda_m^{(2)})^s X_{k,\ell,m}^{(2)}\right|^p\right]\right]^{1/p} \\
&\le \mathbb{E}_{Y^{(1)}}\left[\mathbb{E}_{Y^{(2)}}\left[\left|\sum_{k,\ell,m} (\lambda_\ell^{(1)})^s X_{k,\ell,m}^{(1)} X_{k,\ell,m}^{(2)}\right|^2\right]^{p/2}\right]^{1/p} && \text{(by } G^{(2)}\text{)} \\
&\le \mathbb{E}_{Y^{(2)}}\left[\mathbb{E}_{Y^{(1)}}\left[\left|\sum_{k,\ell,m} (\lambda_\ell^{(1)})^s X_{k,\ell,m}^{(1)} X_{k,\ell,m}^{(2)}\right|^p\right]^{2/p}\right]^{1/2} && \text{(by Minkowski)} \\
&\le \mathbb{E}_{Y^{(2)}}\left[\mathbb{E}_{Y^{(1)}}\left[\left|\sum_{k,\ell,m} X_{k,\ell,m}^{(1)} X_{k,\ell,m}^{(2)}\right|^2\right]\right]^{1/2} = \|X\|_{L^2}. && \text{(by } G^{(1)}\text{)}
\end{aligned}$$

The general case follows from the limit argument. □

3.A.5 Proof of Proposition 3.16

Proof. Let $G = (Y, Q, \lambda)$. Suppose $\widetilde{\deg}_G X < \infty$ and let n be the minimum integer satisfying $X \in \overline{Q_n(Y)}$. Then, by decomposing $X = \sum_{m=0}^n X_m$ with $X_m \in H_m(Y)$, we obtain

$$\|X\|_{L^p} = \left\| T(G)^s \sum_{m=0}^{n} \lambda_m^{-s} X_m \right\|_{L^p} \le \left\| \sum_{m=0}^{n} \lambda_m^{-s} X_m \right\|_{L^2} \le \lambda_m^{-s} \|X\|_{L^2},$$

where we have used the $(2, p; s)$-hypercontractivity in the second inequality. □

3.A.6 Proof of Proposition 3.17

Proof. It suffices to consider X having the decomposition $X = \sum_m X_m$ with $X_m \in H_m(Y)$. Recall that we have assumed that Q_0 is the space of constant functions, so X_0 is a constant. First, suppose $X_0 = 0$. In this case, for $t > s$, we have

$$\begin{aligned} \|T(G)^t X\|_{L^4}^2 = \left\| \sum_{m \ge 1} \lambda_m^t X_m \right\|_{L^4}^2 &\le \left(\sum_{m \ge 1} \lambda_m^{t-s} \lambda_m^s \|X_m\|_{L^4} \right)^2 \\ &\le \left(\sum_{m \ge 1} \lambda_m^{t-s} \|X_m\|_{L^2} \right)^2 \le \left(\sum_{m \ge 1} \lambda_m^{2(t-s)} \right) \|X\|_{L^2}^2. \quad \text{(Cauchy–Schwarz)} \end{aligned}$$

Therefore, when $\sum_{m \ge 1} \lambda_m^{2(t-s)} \le 1/\sqrt{3}$ we have

$$\|T(G)^t X\|_{L^4} \le 3^{-1/4} \|X\|_{L^2} \tag{3.7}$$

for all X satisfying $X_0 = 0$.

In the case $X_0 \ne 0$, we can assume $X_0 = 1$ without loss of generality. Let $W = X - 1$ and $Z = T(G)^t W = T(G)^t X - 1$. Note that $\mathbb{E}[W] = \mathbb{E}[Z] = 0$ holds by the orthogonality. We can explicitly expand the L^4 norm as follows:

$$\begin{aligned} \|T(G)^t X\|_{L^4}^4 &= 1 + 6\mathbb{E}\big[Z^2\big] + 4\mathbb{E}\big[Z^3\big] + \mathbb{E}\big[Z^4\big] \\ &\le 1 + 8\mathbb{E}\big[Z^2\big] + 3\mathbb{E}\big[Z^4\big]. \qquad \text{(AM–GM)} \end{aligned}$$

We also have

$$\|X\|_{L^2}^4 = \mathbb{E}\big[(1+W)^2\big]^2 = (1 + \mathbb{E}\big[W^2\big])^2 = 1 + 2\mathbb{E}\big[W^2\big] + \mathbb{E}\big[W^2\big]^2.$$

So it suffices to show $4\mathbb{E}[Z^2] \le \mathbb{E}[W^2]$ and $3\mathbb{E}[Z^4] \le \mathbb{E}[W^2]^2$, but the latter immediately follows from (3.7). The former holds when $\lambda_1^t \le 1/2$:

$$\mathbb{E}\big[Z^2\big] = \sum_{m \ge 1} \lambda_m^{2t} \mathbb{E}\big[X_m^2\big] \le \lambda_1^{2t} \mathbb{E}\big[W^2\big].$$

Therefore, we have completed the proof. □

3.A.7 Proof of Theorem 3.18

Proof. Let $G = (Y, Q, \lambda)$ and $\mathcal{X}$ be the space in which Y takes values. By truncating Q and λ (i.e., ignoring Q_m with $1/\lambda_m > K$), we can assume that $Q(Y) = \{X \in L^2 \mid \widetilde{\deg}_G X \le K\}$. Then, as $\dim Q < \infty$, we can take a vector-valued measurable function

$$\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_N)^\top : \mathcal{X} \to \mathbb{R}^N$$

such that $(\varphi_i(Y))_{i=1}^N$ is an orthonormal basis of $Q(Y)$. Then, we have

$$\sup_{X \in Q(Y) \setminus \{0\}} \frac{\|X\|_{L^4}}{\|X\|_{L^2}} = \sup_{c \in \mathbb{R}^N \setminus \{0\}} \frac{\|c^\top \boldsymbol{\varphi}(Y)\|_{L^4}}{\|c^\top \boldsymbol{\varphi}(Y)\|_{L^2}} = \sup_{c \in \mathbb{R}^N,\, \|c\|=1} \|c^\top \boldsymbol{\varphi}(Y)\|_{L^4} < \infty,$$

where the right-hand side is the supremum of a continuous functions over a compact domain, and so is indeed finite. Hence, we can apply Proposition 3.17, and there exists a constant $s > 0$ such that

$$\left\|T(G)^t X\right\|_{L^4} \le \|X\|_{L^2}, \qquad X \in Q(Y),$$

because $\lambda_1 < 1$ and $(\lambda_m)_m$ is of finite length now. So $G = (Y, Q, \lambda)$ (with truncation by K) is actually $(2, p; t)$-hypercontractive and it extends to $G^{\otimes d}$ for any d by Theorem 3.15 (note that the truncation does not affect the random variables with $\widetilde{\deg}_{G^{\otimes d}} X \le K$). Then, we finally use Proposition 3.16 to obtain the desired result with $C = K^t$. □

3.A.8 Proof of Proposition 3.25

Proof. Let $f \in L^2(\mu)$ be an eigenfunction with eigenvalue $\lambda \ge 0$ of the integral operator, i.e., it satisfies $\int_{\mathcal{X}} k(x, y) f(y) \,\mathrm{d}\mu(y) = \lambda f(x)$ (assume this equality holds for all x, not just μ-almost all). As $\sum_{\ell=1}^\infty \sigma_\ell < \infty$ and Assumption B is met from the general theory [160], it suffices to show $\lambda \ge 1$ if and only if f is constant. Note that $f = 1$ is an eigenfunction for $\lambda = 1$ by assumption.

Assume $\lambda \ge 1$. Since k is bounded from the assumption, for an $(x_n)_{n=1}^\infty$ converging to x, we have $f(x_n) = \frac{1}{\lambda} \int_{\mathcal{X}} k(x_n, y) f(y) \,\mathrm{d}\mu(y) \to \frac{1}{\lambda} \int_{\mathcal{X}} k(x, y) f(y) \,\mathrm{d}\mu(y) =$

$f(x)$ by the dominated convergence theorem. Thus, f is continuous. Let $F = \max_{x\in\mathcal{X}} f(x)$. If $x^* \in f^{-1}(\{F\})$, then

$$0 = F - f(x^*) = \int_{\mathcal{X}} k(x^*, y)\left(F - \frac{1}{\lambda}f(y)\right) \mathrm{d}\mu(y).$$

As $k(x^*, \cdot)$ is a probability density (recall $k \geq 0$ from the assumption) with respect to μ and $\operatorname{supp}\mu = \mathcal{X}$, we must have $\lambda \leq 1$ and $k(x^*, y) = 0$ for all $y \notin f^{-1}(\{F\})$. Now, it suffices to prove $f^{-1}(\{F\}) = \mathcal{X}$ actually holds when $\lambda = 1$. Let $K = \max_{x,y\in\mathcal{X}} k(x, y)$. By taking an $\varepsilon > 0$ such that $\mu(f^{-1}([F - \varepsilon, F))) \leq 1/(2K)$, we have, for $x \notin f^{-1}(\{F\})$,

$$\begin{aligned} f(x) &= \int_{\mathcal{X}} k(x, y)f(y)\,\mathrm{d}\mu(y) \\ &\leq \int_{f^{-1}((-\infty,F-\varepsilon))} k(x, y)f(y)\,\mathrm{d}\mu(y) + \int_{f^{-1}([F-\varepsilon,F))} k(x, y)f(y)\,\mathrm{d}\mu(y) \\ &\leq (F - \varepsilon)\int_{f^{-1}((-\infty,F-\varepsilon))} k(x, y)\,\mathrm{d}\mu(y) + F\int_{f^{-1}([F-\varepsilon,F))} k(x, y)\,\mathrm{d}\mu(y) \\ &\leq (F - \varepsilon) + \varepsilon\int_{f^{-1}([F-\varepsilon,F))} k(x, y)\,\mathrm{d}\mu(y) \leq (F - \varepsilon) + \frac{\varepsilon}{2} = F - \frac{\varepsilon}{2}. \end{aligned}$$

Therefore, if $f^{-1}(\{F\}) = \mathcal{X}$, f is disconnected (because $\mathcal{X}$ is path-connected), and it is a contradiction. This completes the proof. □

3.A.9 Proof of Proposition 3.26

We first prove the following lemma.

Lemma 3.29. *For $p > 2$, we have $\|\mathcal{K}_0 f\|_{L^p} \leq \|k_0\|_{L^p(\mu\otimes\mu)}\|f\|_{L^2}$ for all $f \in L^2(\mu)$.*

Proof. By Minkowski's integral inequality, we have

$$\begin{aligned}\|\mathcal{K}_0 f\|_{L^p} &= \left(\int_{\mathcal{X}} \left|\int_{\mathcal{X}} k_0(x,y) f(y)\,\mathrm{d}\mu(y)\right|^p \mathrm{d}\mu(x)\right)^{1/p} \\ &\leq \int_{\mathcal{X}} \left(\int_{\mathcal{X}} |k_0(x,y)f(y)|^p \,\mathrm{d}\mu(x)\right)^{1/p} \mathrm{d}\mu(y) \\ &\leq \int_{\mathcal{X}} \left(\int_{\mathcal{X}} |k_0(x,y)|^p \,\mathrm{d}\mu(x)\right)^{1/p} |f(y)|\,\mathrm{d}\mu(y) \\ &\leq \left(\int_{\mathcal{X}} \left(\int_{\mathcal{X}} |k_0(x,y)|^p \,\mathrm{d}\mu(x)\right)^{2/p} \mathrm{d}\mu(y)\right)^{1/2} \|f\|_{L^2} \quad \text{(Cauchy–Schwarz)} \\ &\leq \|k_0\|_{L^p(\mu\otimes\mu)} \|f\|_{L^2}.\end{aligned}$$

□

From this lemma, we have

$$\|e_m\|_{L^p} = \frac{1}{\sigma_m}\|\mathcal{K}_0 e_m\|_{L^p} \leq \frac{\|k_0\|_{L^p(\mu\otimes\mu)}}{\sigma_m}\|e_m\|_{L^2} \tag{3.8}$$

for each $m \geq 2$.

Proof of Proposition 3.26. It suffices to consider the case $\|k_0\|_{L^4(\mu\otimes\mu)} < \infty$. Note that $\lambda_{\ell-1} = \sigma_\ell$ for $\ell = 1, 2, \ldots$ for the GRP $G_{k,\mu}$, so $\lambda_1 = \sigma_2 = \|\mathcal{K}_0\|$.

Let r_0 be the minimum nonnegative number satisfying $\|\mathcal{K}_0\|^{-r_0} \geq \sqrt{3}\,\mathrm{tr}(\mathcal{K}_0)$. Then, for $r := 1 + r_0$, we have

$$\sum_{\ell=2}^{\infty} \sigma_\ell^r \leq \sigma_2^{r_0} \sum_{\ell=2}^{\infty} \sigma_\ell = \|\mathcal{K}\|^{r_0}\,\mathrm{tr}(\mathcal{K}_0) \leq \frac{1}{\sqrt{3}} \tag{3.9}$$

Let s_0 be the minimum nonnegative number satisfying $\|\mathcal{K}_0\|^{-s_0} \geq \|k_0\|_{L^4(\mu\otimes\mu)}$. As $\|\mathcal{K}_0\| \in (0,1)$ from Assumption A, s_0 is well-defined. Then, for $s := 1 + s_0$ and $m \geq 2$, from (3.8), we have

$$\|e_m\|_{L^4} \leq \frac{\|k_0\|_{L^4(\mu\otimes\mu)}}{\sigma_m}\|e_m\|_{L^2} \leq \frac{1}{\sigma_m\|\mathcal{K}_0\|^{s_0}}\|e_m\|_{L^2} \leq \sigma_m^{-1-s_0}\|e_m\|_{L^2}. \tag{3.10}$$

Thus, the condition for s and $t := r + s$ of Proposition 3.17 is satisfied, and so we have the desired conclusion. □

Chapter 4

Positively weighted kernel quadrature via subsampling

In this chapter, we study kernel quadrature rules with convex weights. Our approach combines the spectral properties of the kernel with recombination results about point measures. This results in effective algorithms that construct convex quadrature rules using only access to i.i.d. samples from the underlying measure and evaluation of the kernel and that result in a small worst-case error. In addition to our theoretical results and the benefits resulting from convex weights, our experiments indicate that this construction can compete with the optimal bounds in well-known examples.[1]

4.1 Introduction

The goal of numerical quadrature/cubature is to provide, for a given probability measure μ on a space $\mathcal{X}$, a set of points $x_1, \ldots, x_n \in \mathcal{X}$ and weights $w_1, \ldots, w_n \in \mathbb{R}$ such that

$$\sum_{i=1}^{n} w_i f(x_i) \approx \int_{\mathcal{X}} f(x) \, \mathrm{d}\mu(x) \tag{4.1}$$

holds for a large class of functions $f : \mathcal{X} \to \mathbb{R}$. As already discussed in Section 1.2, kernel quadrature focuses on the case when the function class forms a reproducing kernel Hilbert space (RKHS). What makes kernel quadrature attractive, is that

[1] Code: `https://github.com/satoshi-hayakawa/kernel-quadrature`

the kernel choice provides a simple and flexible way to encode the regularity properties of a function class. Exploiting such regularity properties is essential when the integration domain $\mathcal{X}$ is high-dimensional or the function class is large. Additionally, the domain $\mathcal{X}$ does not have to be Euclidean but can be any topological space that carries a positive definite kernel.

As we have already introduced the notation in Section 1.2, given a set (quadrature) $Q_n = (w_i, x_i)_{i=1}^n$ with $(w_i, x_i) \in \mathbb{R} \times \mathcal{X}$, we also regard it as a quadrature measure $Q_n = \sum_{i=1}^n w_i \delta_{x_i}$ and try to minimize the worst-case error $\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)$, where $\mathcal{H}_k$ denotes the RKHS associated with a positive definite kernel k. If the weights satisfy $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$, we refer to Q_n as a convex quadrature.

Contribution. The primary contribution of this chapter is to leverage recombination (a consequence of Carathéodory's Theorem) with spectral analysis of kernels to construct convex kernel quadrature rules and derive convergence rates. We also provide efficient algorithms that compute these quadrature rules; they only need access to i.i.d. samples from μ and the evaluation of the kernel k. See Table 4.1 for a comparison with other kernel quadrature constructions.

The table is written by using σ_n and r_n, which represents a sort of decay of the kernel with respect to μ. Typical regimes are $\sigma_n \sim n^{-\beta}$ (e.g. Sobolev) or $\sigma_n \sim \exp(-\gamma n)$ (e.g. Gaussian) depending on the 'smoothness' of the kernel [e.g., 49, 6] (see also Section 4.B.3), and in such regimes (with $\beta \geq 2$ or $\gamma > 0$), σ_n or $r_n (\lesssim n\sigma_n)$ provide faster rates than $\mathrm{wce}^2 \sim 1/n$ of the usual Monte Carlo rate. For more examples including multivariate Sobolev spaces, see Bach [6, Section 2.3].

Limitation. Our proposed methods are based on either Mercer or Nyström approximation. Though our Mercer-based methods result in strong theoretical bounds, they require the knowledge of Mercer decomposition like [6, 16, 17], which is not available for general (k, μ). Our Nyström-based methods apply to much more general situations and outperform existing methods in experiments, but the $n/\sqrt{\ell}$ term makes their theoretical bound far from competitive. Further study is needed to bridge the gap between theory and empirical results. This point is addressed in the next chapter under additional assumptions on spectral decay.

Table 4.1: Comparison on n-point kernel quadrature rules. We are omitting the $\mathcal{O}$ notation throughout the table. Note that the assumption under which the theoretical guarantee holds varies from method to method, and this table displays just a representative bound derived in the cited references. Here are remarks on the notation. (1) σ_m is the m-th eigenvalue of the integral operator $\mathcal{K}$, and $r_m = \sum_{i=m}^{\infty} \sigma_i$. (2) The symbols in the first line respectively mean C: convex, M: *not* using the knowledge of Mercer decomposition, and E: *not* using the knowledge of expectations such as $\int_{\mathcal{X}} k(x, y) \, \mathrm{d}\mu(y)$. (3) The ($m$: global optimization) is indicating the cost of globally optimizing a function whose evaluation costs $\Theta(m)$. (4) N (N_φ) refers to the size of the candidate set from which we subsample, s is a batch size in the batch herding, and ℓ is another parameter for the Nyström-based methods. (†) Mercer/Nyström are the algorithms based on random convex hulls, see Section 4.2.4 and Appendix 4.D. (‡) M./N. + empirical are the algorithms discussed in the main text. References for the itemized methods are as follows: Herding [30, 7], Batch herding [164], SBQ [78], Leveraged [6], DPP [16, 15], CVS [17], RPCholesky [46], and KT++ [44, 45, 152]. We may have better bounds for specific cases such as i.i.d. sampling for Sobolev spaces with uniform measure [96], but such methods are not included.

Method	Bound of squared wce	Computational complexity	C	M	E
Herding	$1/n$	$n \cdot (n\text{: global optimization})$	✓	✓	
Batch herding	$(s \log(n/s) \log N)/n + 1/N$	$n^2 N^s$	✓	✓	
SBQ	Not found	$n \cdot (n^2\text{: global optimization})$		✓	
Leveraged	σ_m, $m \lesssim n/\log n$	Unavailable			
DPP	r_{n+1}	$n^3 \cdot$ (rejection sampling)			
CVS	σ_{n+1} (with assumptions)	Unavailable		✓	
RPCholesky	$\frac{r_{m+1}}{m}$, $n \gtrsim m \log(\frac{m\sigma_1}{r_{m+1}})$	$n \cdot$ (rejection sampling)		✓	
KT++	$(1/n^2 + 1/N) \operatorname{polylog}(N)$	$N \log^3 N$	✓	✓	✓
Ours:					
Mercer†	r_n	$nN_\varphi + C(n, N_\varphi)$	✓		
M. + empirical‡	$r_n + \frac{1}{N}$	$nN + n^3 \log(N/n)$	✓		✓
Nyström†	$n\sigma_n + r_{n+1} + \frac{n}{\sqrt{\ell}}$	$n\ell N_\varphi + n\ell^2 + C(n, N_\varphi)$	✓	✓	
N. + empirical‡	$n\sigma_n + r_{n+1} + \frac{n}{\sqrt{\ell}} + \frac{1}{N}$	$n\ell N + n\ell^2 + n^3 \log(N/n)$	✓	✓	✓

Why convex weights? There are several reasons why convex weights are preferable: (i) **Positive integral operator:** Kernel quadrature provides an approximation of the integration operator $f \mapsto I(f) = \int f(x) \, \mathrm{d}\mu(x)$. Hence, a natural requirement is to preserve some basic properties of this operator and positive weights preserve the positivity of this operator. (ii) **Uniform estimates and robustness:** In applications, the RKHS $\mathcal{H}_k$ may be misspecified and, if a quadrature rule with possibly negative weights and large total variation ($\sum_{i=1}^n |w_i|$ in (4.1)) is applied to a function $f \notin \mathcal{H}_k$, the approximation error (4.1) can get large; in contrast, a simple estimate shows that convex weights give uniform bounds, see Appendix 4.B.4. (iii) **Iteration:** Consider the m-fold product of quadrature formulas for approximating $\mu^{\otimes m}$ on $\mathcal{X}^m$. This is a common construction for multidimensional quadrature formulas (e.g., for polynomials) from one-dimensional formulas [163] or numerics for stochastic differential equations [110]. In doing so, working with a probability measure is strongly preferred, since otherwise, the total variation of their m-fold product gets exponentially large as m increases ($\|Q^{\otimes m}\|_{\mathrm{TV}} = \|Q\|_{\mathrm{TV}}^m$ for a quadrature Q, where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm).

Related literature. Roughly speaking, there have been two approaches to kernel-based quadrature formulas: kernel herding and random sampling. In kernel herding or its variants, the points $(x_i)_{i=1}^n$ are found iteratively, typically based on the Frank–Wolfe gradient descent algorithm [30, 7, 78].

In the random sampling approach, $(x_i)_{i=1}^n$ are sampled and subsequently, the weights are optimized. Generically, this results only in a signed measure μ^Q but not a probability measure. Bach [6] and Belhadji et al. [16] use the eigenvalues and the eigenfunctions of the integral operator $\mathcal{K} : f \mapsto \int_{\mathcal{X}} k(\cdot, y) f(y) \,\mathrm{d}\mu(y)$ to obtain a Mercer-type decomposition of k [160]. Bach [6] then uses the eigenvalues and eigenfunctions of $\mathcal{K}$ to define an optimized measure from which the points $(x_i)_{i=1}^n$ are i.i.d. sampled. This achieves a near-optimal rate, but the exact sampling from this measure is usually unavailable, although, for special cases, it can be done efficiently. In contrast, Belhadji et al. [16] proposes non-independent sampling based on determinantal point processes (DPPs [76]). These two papers also treat the more general quadrature problem that includes a weight function $g \in L^2(\mu)$, i.e., approximating $\int_{\mathcal{X}} f(x)g(x) \,\mathrm{d}\mu(x)$ for $f \in \mathcal{H}_k \subset L^2(\mu)$, which we do not discuss in this chapter. Another recently introduced method is kernel thinning [44, 45], which aims at efficient compression of empirical measures that can be obtained by sampling like our '+ empirical' methods. Its acceleration [152] makes it competitive in terms of compressing $N \sim n^2$ points ('KT++' in Table 4.1).

Finally, we emphasize that the kernel quadrature literature is vast, and the distinction between herding and sampling is only a rough dichotomy, see e.g. [39, 106, 22, 82, 129, 85, 83, 156]. Beyond kernel quadrature, our algorithms can also contribute to the density estimation approach in [170] which relies on recombination based on Fourier features although we do not pursue this further here.

Outline. Section 4.2 contains our main theoretical and methodological contribution. Section 4.3 provides numerical experiments on common benchmarks. The Appendix contains several extensions of our main result, proofs, and further experiments and benchmarks.

4.2 Main result

Assume we are given a set[2] of $n-1$ functions $\varphi_1, \dots, \varphi_{n-1} : \mathcal{X} \to \mathbb{R}$ such that their linear combinations well approximate functions in $\mathcal{H}_k$. Then our kernel quadrature problem reduces to the construction of an n-point discrete probability measure $\mu^{Q_n} = \sum_{i=1}^n w_i \delta_{x_i}$ such that

$$\int_{\mathcal{X}} \varphi_i(x) \, \mathrm{d}\mu^{Q_n}(x) = \int_{\mathcal{X}} \varphi_i(x) \, \mathrm{d}\mu(x) \quad \text{for every } i = 1, \dots, n-1. \tag{4.2}$$

A simple way to *approximately* construct this μ^{Q_n} is to first, sample $N \gg n$ points, $(y_i)_{i=1}^N$, from μ such that their empirical measure, $\widetilde{\mu}_N = \frac{1}{N} \sum_{i=1}^N \delta_{y_i}$, is a good approximation to μ in the sense that $\int \varphi_i \, \mathrm{d}\widetilde{\mu}_N \approx \int \varphi_i \, \mathrm{d}\mu$ for $i = 1, \dots, n-1$, and secondly, apply a so-called recombination algorithm (Remark 4.1) that takes as input $(y_i)_{i=1}^N$ and n functions $\varphi_1, \dots, \varphi_{n-1}$ and outputs a measure $\mu^{Q_n} = \sum w_i \delta_{x_i}$ by selecting a subset $(x_i)_{i=1}^n$ of the points $(y_i)_{i=1}^N$ and giving them weights $(w_i)_{i=1}^n$ such that μ^{Q_n} is a probability measure that satisfies the equation (4.2) with μ replaced by $\widetilde{\mu}_N$.

[2] The number $n-1$ stems from Carathéodory's theorem, Remark 4.1, and leads to an n point quadrature rule.

The challenging parts of this approach are (i) to construct functions $\varphi_1, \dots, \varphi_{n-1}$ that approximately span the RKHS $\mathcal{H}_k$ for a small n; (ii) to arrive at good quantitative bounds despite the (probabilistic) sampling error resulting from the use of the empirical measure $\widetilde{\mu}_N$, and the function approximation error via $\varphi_1, \dots, \varphi_{n-1}$. To address (i) we look for functions such that

$$k(x, y) \approx k_0(x, y) := \sum_{i=1}^{n-1} c_i \varphi_i(x) \varphi_i(y) \tag{4.3}$$

with some $c_i \geq 0$. Two classic ways to do this are the Mercer and Nyström approximations. The remaining, item (ii) is our main contribution. Theorem 4.1 shows that the worst-case error, (4.4), is controlled by the sum of two terms: the first term stems from the kernel approximation (4.3), the second term stems from the empirical measure.

Theorem 4.1. *Let μ be a Borel probability measure on $\mathcal{X}$ and k a positive definite kernel on $\mathcal{X}$ such that $\int_{\mathcal{X}} k(x, x) \,\mathrm{d}\mu(x) < \infty$. Further, let n be a positive integer and assume k_0 is a positive definite kernel on $\mathcal{X}$ such that*

1. $k - k_0$ is a positive definite kernel on $\mathcal{X}$, and 2. $\dim \mathcal{H}_{k_0} < n$.

There exists a function KQuad *such that if D_N is a set of N i.i.d. samples from μ, then $Q_n = \mathrm{KQuad}(D_N)$ is a random n-point convex quadrature that satisfies*

$$\mathbb{E}_{D_N}\left[\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2\right] \leq 8 \int_{\mathcal{X}} (k(x, x) - k_0(x, x)) \,\mathrm{d}\mu(x) + \frac{2c_{k,\mu}}{N}. \tag{4.4}$$

where $c_{k,\mu} := \int_{\mathcal{X}} k(x, x) \,\mathrm{d}\mu(x) - \iint_{\mathcal{X}\times\mathcal{X}} k(x, y) \,\mathrm{d}\mu(x) \,\mathrm{d}\mu(y)$.

Moreover, the support of Q_n is a subset of D_N and given functions $\varphi_1, \dots, \varphi_{n-1} \in L^1(\mu)$ with $\mathcal{H}_{k_0} \subset \mathrm{span}\{\varphi_1, \dots, \varphi_{n-1}\}$, $Q_n = \mathrm{KQuad}(D_N)$ can be computed with Algorithm 4.1 in $\mathcal{O}(nN + n^3 \log(N/n))$ computational steps.

The function KQuad is deterministic but since D_N is random, the resulting quadrature rule Q_n is random, hence also the resulting worst-case error $\mathrm{wce}(Q; \mathcal{H}_k, \mu)$ and the expectation in (4.4) denotes the expectation over the N samples in D_N. The theoretical part of Theorem 4.1 follows from more general results that we present and prove in the Appendix: Theorem 4.7 proves the inequality, essentially by comparing $\mathcal{H}_k$ with $\mathcal{H}_{k_0}$; Theorem 4.9 proves the existence. The algorithmic part of Theorem4.1 is discussed in Section 4.2.1 below. Theorem 4.1 covers our two main examples for the construction of k_0, resp. the choice of $\varphi_1, \dots, \varphi_{n-1}$, and for which the error estimate gets quite explicit: the Mercer approximation, see Section 4.2.2, and the Nyström approximation, see Section 4.2.3. The former requires some knowledge about the spectrum of the kernel which is, however, known for many popular kernels; the latter works in full generality but yields worse theoretical guarantees for the convergence rate. Finally, we emphasize that N and n in Theorem 4.1 can be chosen independently and we will see that from a computational point the choice $N \sim n^2$ is preferable in which case (4.4) is a faster rate than Monte Carlo, see also Table 4.1.

4.2.1 Algorithm

Algorithm 4.1 Kernel quadrature with convex weights via recombination KQuad

```
Input: A positive definite kernel $k$ on $\mathcal{X}$, a probability measure $\mu$ on $\mathcal{X}$, integers
    $N \geq n \geq 1$, another kernel $k_0$, functions $\varphi_1, \dots, \varphi_{n-1}$ on $\mathcal{X}$ with $\mathcal{H}_{k_0} \subset$
    $\mathrm{span}\{\varphi_1, \dots, \varphi_{n-1}\}$ and a set $D_N$ of $N$ i.i.d. samples from $\mu$.
Output: A set $Q_n \coloneqq \{(w_i, x_i) \mid i = 1, \dots, n\} \subset \mathbb{R} \times \mathcal{X}$ with $w_i \geq 0$, $\sum_{i=1}^n w_i = 1$
  Apply a Recombination Algorithm (Remark 4.1) with $\boldsymbol{\psi} =$
    $(\varphi_1, \dots, \varphi_{n-1}, k_{1,\mathrm{diag}})^\top$, to the empirical measure $\frac{1}{N}\sum_{y \in D_N} \delta_y$ to obtain points
    $\{\widetilde{x}_1, \dots, \widetilde{x}_{n+1}\} \subset D_N$ and weights $\boldsymbol{v} = (v_1, \dots, v_{n+1})^\top \geq \mathbf{0}$ that satisfy $\mathbf{1}^\top \boldsymbol{v} = 1$
    and $\boldsymbol{\psi}(\widetilde{\boldsymbol{x}})\boldsymbol{v} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\psi}(\widetilde{x}_i)$, where $\boldsymbol{\psi}(\widetilde{\boldsymbol{x}}) = [\boldsymbol{\psi}(\widetilde{x}_1), \dots, \boldsymbol{\psi}(\widetilde{x}_{n+1})] \in \mathbb{R}^{n \times (n+1)}$.
  Apply SVD with the matrix $A = [\varphi_{i-1}(y_j)]_{ij} \in \mathbb{R}^{n \times (n+1)}$ with $\varphi_0 = 1$ to find a
    nonzero vector $\boldsymbol{u} \in \mathbb{R}^{n+1}$ such that $A\boldsymbol{u} = \mathbf{0}$ and $k_{1,\mathrm{diag}}(\widetilde{\boldsymbol{x}})^\top \boldsymbol{u} \geq 0$
  Compute the smallest $\alpha \geq 0$ such that $\boldsymbol{v} - \alpha\boldsymbol{u} \geq \mathbf{0}$ and $v_j - \alpha u_j = 0$ for a $j$
  Return $(w_i)_{i=1}^n \leftarrow (v_k - \alpha u_k)_{k \in I}$ and $(x_i)_{i=1}^n \leftarrow (\widetilde{x}_k)_{k \in I}$, where $I = \{1, \dots, n +$
    $1\} \setminus \{j\}$
```

Suppose we are given k_0 and $\varphi_1, \dots, \varphi_{n-1} \in L^1(\mu)$ with $\mathcal{H}_{k_0} \subset \mathrm{span}\{\varphi_1, \dots, \varphi_{n-1}\}$, and also N independent samples from μ denoted by $D_N = (y_1, \dots, y_N)$. Theorem 4.7 in the Appendix shows that if we construct a convex quadrature $Q_n = (w_i, x_i)_{i=1}^n$ satisfying

$$\sum_{i=1}^n w_i \boldsymbol{\varphi}(x_i) = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\varphi}(y_i), \qquad \sum_{i=1}^n w_i k_{1,\mathrm{diag}}(x_i) \leq \frac{1}{N}\sum_{i=1}^N k_{1,\mathrm{diag}}(y_i), \tag{4.5}$$

where $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_{n-1})^\top$ and $k_{1,\mathrm{diag}}(x) = k(x,x) - k_0(x,x)$, it satisfies the bound (4.4). For this problem, we can use the so-called *recombination* algorithms:

Remark 4.1 (Recombination)**.** *Given $d-1$ functions (called test functions) and a probability measure supported on $N > d$ points, there exists a probability measure supported on a subset of d points that gives the same mean to these $d-1$ functions. This follows from Carathéodory's theorem and is known as recombination. Efficient deterministic [103, 111, 163] as well as randomized [34] algorithms exist to compute the new probability measure supported on d points; e.g. deterministic algorithms perform the recombination, step 1, in $\mathcal{O}(c_\varphi N + d^3 \log(N/d))$ time, where c_φ is the cost of computing all the test functions at one sample. If each function evaluation is in constant time, $c_\varphi = \mathcal{O}(d)$.*

Let us briefly provide the intuition behind the deterministic recombination algorithms. We can solve the problem of "reducing (weighted) $2d$ points to d points in $\mathbb{R}^d$ while keeping the barycenter" by using linear programming or a variant of it. If we apply this to $2d$ points each given by a barycenter of approximately $\frac{N}{2d}$ points, we can reduce the original problem of size N to a problem of size $d \cdot \frac{N}{2d} = \frac{N}{2}$. By repeating this procedure $\log_2(\frac{N}{d})$ times we obtain the desired measure.

Although the recombination introduced here only treats the equality constraints in (4.5) we can satisfy the remaining constraints just with n points by modifying it. This is done in Algorithm 4.1 which works as follows: First, via recombination, find an $(n+1)$-point convex quadrature $R_{n+1} = (v_i, y_i)_{i=1}^{n+1}$ that exactly integrates functions $\varphi_1, \dots, \varphi_{n-1}, k_{1,\mathrm{diag}}$ with regard to the empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{y_i}$. Second, to reduce one point, find a direction ($-\boldsymbol{u}$ in the algorithm) in the space of weights on $(\widetilde{x}_i)_{i=1}^{n+1}$ that does not change the integrals of $\varphi_1, \dots, \varphi_{n-1}$ and the constant function 1, and does not increase the integral of $k_{1,\mathrm{diag}}$. Finally, move the weight from $\boldsymbol{v}$ to the above direction until an entry becomes zero, at $\boldsymbol{v} - \alpha\boldsymbol{u}$. Such an $\alpha \geq 0$ exists, as $\boldsymbol{u}$ must have a positive entry since it is a nonzero vector whose entries sum up to one. Now we have a convex weight vector with at most n nonzero entries, so it outputs the desired quadrature satisfying (4.5).

4.2.2 Mercer approximation

In this section and Section 4.2.3, we assume that k has a pointwise convergent Mercer decomposition $k(x,y) = \sum_{m=1}^\infty \sigma_m e_m(x) e_m(y)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ and $(e_m)_{m=1}^\infty \subset L^2(\mu)$ being orthonormal [160]. If we let $\mathcal{K}$ be the integral operator $L^2(\mu) \to L^2(\mu)$ given by $f \mapsto \int_{\mathcal{X}} k(\cdot, y) f(y) \, \mathrm{d}\mu(y)$, then $(\sigma_m, e_m)_{m=1}^\infty$ are the eigenpairs of this operator.

The first choice of the approximate kernel k_0 is just the truncation of Mercer decomposition.

Corollary 4.2. *Theorem 4.1 applied with $k_0(x,y) = \sum_{m=1}^{n-1} \sigma_m e_m(x) e_m(y)$ yields a random convex quadrature rule Q_n such that*

$$\mathbb{E}_{D_N}\left[\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2\right] \leq 8 \sum_{m=n}^\infty \sigma_m + \frac{2c_{k,\mu}}{N}. \tag{4.6}$$

Proof. It suffices to prove the result under the assumption $\int_{\mathcal{X}} k(x,x) \, \mathrm{d}\mu(x) = \sum_{m=1}^\infty \sigma_m < \infty$, as otherwise the right-hand side of (4.6) is infinity.

For $k_1 := k - k_0$, we have that $k_1(x,y) = \sum_{m=n}^\infty \sigma_m e_m(x) e_m(y)$ and it is the inner product of $\Phi(x) := (\sqrt{\sigma_m} e_m(x))_{m=n}^\infty$ and $\Phi(y)$ in $\ell^2(\{n, n+1, \ldots\})$ and so positive definite. Thus k and k_0 satisfies the assumption of Theorem 4.1, and $\int_{\mathcal{X}} k_1(x,x) \, \mathrm{d}\mu(x) = \sum_{m=n}^\infty \sigma_m$ applied to (4.4) yields the desired inequality. □

4.2.3 Nyström approximation

Although the Nyström method [179, 43, 97] is primarily used for approximating a large Gram matrix by a low-rank matrix, it can also be used for directly approximating the kernel function itself and this is how we use it. Given a set of ℓ points $Z = (z_i)_{i=1}^\ell \subset \mathcal{X}$, the vanilla Nyström approximation of $k(x,y)$ is given by

$$k(x,y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} \approx \langle P_Z k(\cdot, x), P_Z k(\cdot, y) \rangle_{\mathcal{H}_k} =: k^Z(x,y), \tag{4.7}$$

where $P_Z : \mathcal{H}_k \to \mathcal{H}_k$ is a projection operator onto $\mathrm{span}\{k(\cdot, z_i)\}_{i=1}^\ell$. In matrix notation, we have

$$k^Z(x,y) = k(x,Z) k(Z,Z)^+ k(Z,y) := [k(x,z_1), \ldots, k(x,z_\ell)] k(Z,Z)^+ \begin{bmatrix} k(z_1,y) \\ \vdots \\ k(z_\ell, y) \end{bmatrix}, \tag{4.8}$$

where $k(Z, Z) := (k(z_i, z_j))_{i,j=1}^{\ell}$ is the Gram matrix for Z and $k(Z, Z)^+$ denotes its Moore–Penrose inverse. We discuss the equivalence between (4.7) and (4.8) in Appendix 4.B.5. As k^Z is an ℓ-dimensional kernel, there exists an $(\ell + 1)$-point quadrature formula that exactly integrates functions in $\mathcal{H}_{k^Z}$. For a quadrature formula, exactly integrating all the functions in $\mathcal{H}_{k^Z}$ is indeed equivalent to exactly integrating $k(z_i, \cdot)$ for all $1 \leq i \leq \ell$, as long as the Gram matrix $k(Z, Z)$ is nonsingular. Proposition 4.8 in the Appendix provides a bound for the associated worst-case error. From this viewpoint, the Nyström approximation offers a natural set of test functions.

The Nyström method has a generalization with a low-rank approximation of $k(Z, Z)$. Concretely, let $k(Z, Z)_s$ be the best rank-s approximation of $k(Z, Z)$ (given by eigendecomposition), and we define the following s-dimensional kernel:

$$k_s^Z(x, y) := k(x, Z)k(Z, Z)_s^+ k(Z, y). \tag{4.9}$$

Let $k(Z, Z) = U\Lambda U^\top$ be the eigendecomposition of $k(Z, Z)$, where $U = [u_1, \dots, u_\ell] \in \mathbb{R}^{\ell\times\ell}$ is a real orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_\ell)$ with $\lambda_1 \geq \cdots \geq \lambda_\ell \geq 0$. Then, if $\lambda_s > 0$ we have

$$k_s^Z(x, y) = \sum_{i=1}^{s} \frac{1}{\lambda_i}(u_i^\top k(Z, x))(u_i^\top k(Z, y)). \tag{4.10}$$

So we can use functions $u_i^\top k(Z, \cdot)$ $(i = 1, \dots, s)$ as test functions, which is chosen from a larger dimensional space $\mathrm{span}\{k(z_i, \cdot)\}_{i=1}^{\ell}$. Although closer to the original usage of the Nystöm method is to obtain $u_i^\top k(Z, \cdot)$ as an approximation of i-th eigenfunction of the integral operator $\mathcal{K}$ with Z appropriately chosen with respect to μ, we have adopted an explanation suitable for the machine learning literature [43, 97].

The following is a continuous analogue of Kumar et al. [97, Theorem 2] showing the effectiveness of the Nyström method. See also Jin et al. [80] for an analysis specific to the case $s = \ell$.

Theorem 4.3. *Let $s \leq \ell$ be positive integers and $\delta > 0$. Let Z be an ℓ-point independent sample from μ. If we define the integral operator $\mathcal{K}_s^Z : L^2(\mu) \to L^2(\mu)$*

by $f \mapsto \int_{\mathcal{X}} k_s^Z(\cdot, y) f(y) \, \mathrm{d}\mu(y)$, then we have, with probability at least $1-\delta$, in terms of the operator norm,

$$\|\mathcal{K}_s^Z - \mathcal{K}\| \leq \sigma_{s+1} + \frac{2 \sup_{x \in \mathcal{X}} k(x,x)}{\sqrt{\ell}} \left(1 + \sqrt{2 \log \frac{1}{\delta}} \right). \tag{4.11}$$

The proof is given in Appendix 4.C.5. By using this estimate, we obtain the following guarantee for the random convex quadrature given by Algorithm 4.1 and the Nyström approximation.

Corollary 4.4. *Let D_N be N-point independent sample from μ and let Z be an ℓ-point independent sample from μ. Theorem 4.1 applied with the Nyström approximation $k_0 = k_{n-1}^Z$ yields an random n-point convex quadrature rule Q_n such that, with probability at least $1-\delta$ and $k_{\max} := \sup_{x \in \mathcal{X}} k(x,x)$,*

$$\mathbb{E}_{D_N}\left[\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \,\middle|\, Z\right] \leq 8\left(n\sigma_n + \sum_{m>n} \sigma_m \right) + \frac{16(n-1)k_{\max}}{\sqrt{\ell}} \left(1 + \sqrt{2\log\frac{1}{\delta}} \right) + \frac{2c_{k,\mu}}{N}.$$

Proof. From (4.10), $k^Z(x,y) - k_{n-1}^Z(x,y) = \sum_{i=n}^{\ell} \lambda_i^{-1} (u_i^\top k(Z,x))(u_i^\top k(Z,y))$ (ignore the terms with $\lambda_i = 0$ if necessary), and it is thus positive (semi)definite. If we define $P_Z^\perp : \mathcal{H}_k \to \mathcal{H}_k$ as the projection operator onto the orthogonal complement of $\mathrm{span}\{k(\cdot, z_i)\}_{i=1}^{\ell}$, then, from (4.7), we also have $k(x,y) - k^Z(x,y) = \langle P_Z^\perp k(\cdot,x), P_Z^\perp k(\cdot, y)\rangle_{\mathcal{H}_k}$. Now $k - k^Z$ is also positive definite since we have $\boldsymbol{a}^\top (k(X,X) - k^Z(X,X))\boldsymbol{a} = \sum_{i,j=1}^{M} a_i a_j \langle P_Z^\perp k(\cdot, x_i), P_Z^\perp k(\cdot, x_j)\rangle_{\mathcal{H}_k} = \|\sum_{i=1}^M a_i P_Z^\perp k(\cdot, x_i)\|_{\mathcal{H}_k}^2$ for any $M > 0$, $\boldsymbol{a} = (a_i)_{i=1}^M \in \mathbb{R}^M$ and $X = (x_i)_{i=1}^M \in \mathcal{X}^M$. In particular, $k - k_{n-1}^Z = (k - k^Z) + (k^Z - k_{n-1}^Z)$ is positive definite. Also, it suffices to prove the result when $\sum_{m=1}^\infty \sigma_m < \infty$, so we can now apply Theorem 4.1.

For $k_1 := k - k_{n-1}^Z$, we prove the inequality $\int_{\mathcal{X}} k_1(x,x) \, \mathrm{d}\mu(x) = \sum_{m=1}^\infty \langle e_m, (\mathcal{K} - \mathcal{K}_s^Z) e_m \rangle_{L^2} \leq (n-1)\|\mathcal{K} - \mathcal{K}_s^Z\| + \sum_{m \geq n} \sigma_m$ (see (4.30) in Appendix 4.D.2 for details), and the desired inequality follows by combining Theorem 4.1 and Theorem 4.3 (i.e., (4.4) and (4.11)). □

Remark 4.2. *Algorithm 4.1 with the Nystöm approximation can be decomposed into two parts: (a) Nystöm approximation by truncated singular value decomposition (SVD) (the first $n-1$ eigenvectors from an ℓ-point sample), (b) Recombination from an N-point empirical measure. The complexity of (a) is $\mathcal{O}(n\ell^2)$, and* *it can also be approximated by randomized SVD in $\mathcal{O}(n^2\ell + \ell^2 \log n)$ [63]. The cost of part (b) is $\mathcal{O}(n\ell N + n^3 \log(N/n))$, where $n\ell N$ stems from the evaluation of $k_{1,\mathrm{diag}}$ for all N sampling points. If we do not impose the inequality constraint regarding $k_{1,\mathrm{diag}}$, which still works well in practice, the cost of part (b) becomes $\mathcal{O}(\ell N + n^2\ell \log(N/n))$, by using the trick $\frac{1}{N}\sum_{i=1}^N U_{n-1}^\top k(Z, y_i) = U_{n-1}^\top \frac{1}{N}\sum_{i=1}^N k(Z, y_i)$, where $U_{n-1} = [u_1, \dots, u_{n-1}] \in \mathbb{R}^{\ell\times(n-1)}$ is a truncation of the matrix that appears in the Nyström approximation (4.9,4.10). So the overall complexity is given by $\mathcal{O}(n\ell N + n\ell^2 + n^3 \log(N/n))$ while an approximate algorithm (randomized SVD, without the inequality constraint) runs in $\mathcal{O}(\ell N + \ell^2 \log n + n^2\ell \log(N/n))$.*

4.2.4 Kernel quadrature using expectations of test functions

Algorithm 4.1 and the bound (4.4) can be generally applicable once we obtain a low-rank approximation k_0 as we have seen in Section 4.2.2 and 4.2.3. However, since by construction, we start by reducing the empirical measure given by D_N, it is inevitable to have the $\Omega(1/N)$ term in the error estimate and performance. We can avoid this limitation by exploiting additional knowledge of expectations.

Let k_0 and k_1 be positive definite kernels with $k = k_0 + k_1$. Let $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_{n-1})^\top$ be the vector of test functions that spans $\mathcal{H}_{k_0}$. When we know the expectations of them, i.e., $\int_{\mathcal{X}} \boldsymbol{\varphi}(x)\,\mathrm{d}\mu(x)$, we can actually construct a convex quadrature $Q_n = (w_i, x_i)_{i=1}^n$ satisfying

$$\sum_{i=1}^n w_i \boldsymbol{\varphi}(x_i) = \int_{\mathcal{X}} \boldsymbol{\varphi}(x)\,\mathrm{d}\mu(x), \qquad \sum_{i=1}^n w_i k_1(x_i, x_i) \le \int_{\mathcal{X}} k_1(x, x)\,\mathrm{d}\mu(x) \tag{4.12}$$

with a positive probability by an algorithm based on random convex hulls (Appendix 4.D, Algorithm 4.2). For this Q_n, we have the following theoretical guarantee (see Theorem 4.6 in Appendix 4.B):

Theorem 4.5. *If a convex quadrature Q_n satisfies the condition* (4.12)*, we have*

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \le 4 \int_{\mathcal{X}} k_1(x, x)\,\mathrm{d}\mu(x).$$

If k_0 is given the Mercer/Nyström approximations, we immediately have the following guarantees; they correspond to **Mercer** and **Nyström** in Table 4.1. See also Theorem 4.14 and 4.16 for details.

- If $k_0(x, y) = \sum_{m=1}^{n-1} \sigma_m e_m(x) e_m(y)$ is given by the Mercer approximation, for a convex quadrature Q_n satisfying (4.12), we have

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \le 4 \sum_{m=n}^{\infty} \sigma_m.$$

- Let $k_0 = k_{n-1}^Z$ be given by the Nyström approximation (4.9) with Z being an ℓ-point independent sample from μ (with $\ell > n$). Then, for a convex quadrature Q_n satisfying (4.12), with probability at least $1 - \delta$ (with respect to Z) and $k_{\max} := \sup_{x \in \mathcal{X}} k(x, x)$, we have

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \le 4 \left(n\sigma_n + \sum_{m>n} \sigma_m \right) + \frac{8(n-1)k_{\max}}{\sqrt{\ell}} \left(1 + \sqrt{2 \log \frac{1}{\delta}} \right).$$

4.3 Numerical experiments

In this section, we compare our methods with several existing methods. In all the experiments, we used the setting where we can compute $\int_{\mathcal{X}} k(x, y) \, \mathrm{d}\mu(y)$ for $x \in \mathcal{X}$ and $\iint_{\mathcal{X} \times \mathcal{X}} k(x, y) \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(y)$ since then we can evaluate the worst-case error of quadrature formulas explicitly. Indeed, if a quadrature formula Q_n is given by points $X = (x_i)_{i=1}^n$ and weights $\boldsymbol{w} = (w_i)_{i=1}^n$, then we have

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 = \boldsymbol{w}^\top k(X, X) \boldsymbol{w} - 2\mathbb{E}_y[\boldsymbol{w}^\top k(X, y)] + \mathbb{E}_{y,y'}[k(y, y')] \tag{4.13}$$

for independent $y, y' \sim \mu$ under $\int_{\mathcal{X}} \sqrt{k(x, x)} \, \mathrm{d}\mu(x) < \infty$, which is a special case of (1.2). An essential remark shown in Huszár and Duvenaud [78] is that the Bayesian quadrature [131] with covariance kernel k given observation at points $(x_i)_{i=1}^n$ (automatically) estimates the integral as $\sum_{i=1}^n w_i f(x_i)$ with $(w_i)_{i=1}^n$ minimizing the above expression. Once given points $(x_i)_{i=1}^n$ and additional knowledge of expectations, we can compute the optimal weights $(w_i)_{i=1}^n$ by solving a convex quadratic programming (CQP), either without any restrictions or with the condition that $(w_i)_{i=1}^n$ is convex. Although the former can be solved by matrix inversion, we have used the optimizer Gurobi[3] for both CQPs to avoid numerical instability. For the

[3] Version 9.1.2, `https://www.gurobi.com/`

recombination part, we have modified the Python library by Cosentino et al. [34] implementing the algorithm of [163].

Our theoretical bounds are close to optimal in classic examples and we see that the algorithm even outperforms the theory in practice, especially in Section 4.3.1. We also execute a measure reduction of a large discrete measure in terms of Gaussian RKHS and our methods show a fast convergence rate in two ML datasets in Section 4.3.2. [4]

4.3.1 Periodic Sobolev spaces with uniform measure

For a positive integer r, consider the Sobolev space of functions on $[0, 1]$ endowed with the norm $\|f\|^2 = (\int_0^1 f(x)\,\mathrm{d}x)^2 + (2\pi)^{2r} \int_0^1 f^{(r)}(x)^2\,\mathrm{d}x$, where f and its derivatives $f^{(1)}, \dots f^{(r)}$ are periodic (i.e., $f(0) = f(1)$ and so forth). This function space can be identified as the RKHS of the kernel

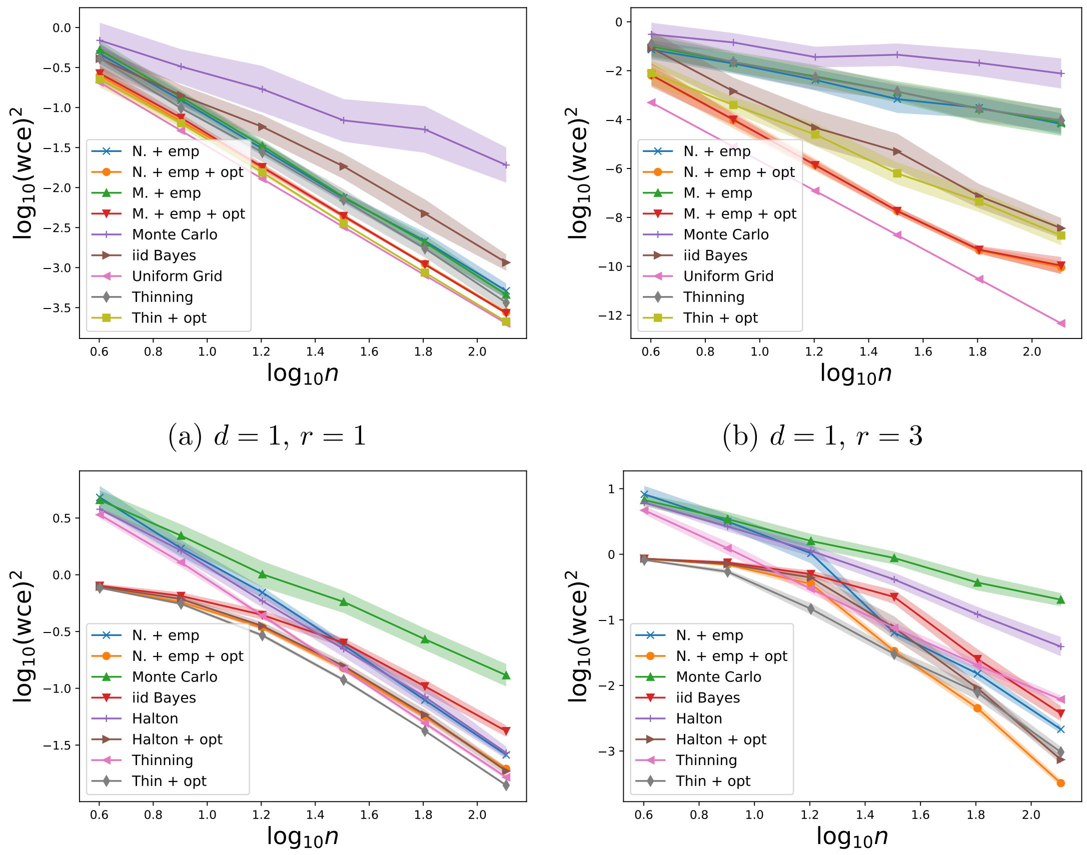
$$k_r(x, y) = 1 + \frac{(-1)^{r-1}(2\pi)^{2r}}{(2r)!} B_{2r}(|x - y|)$$

for $x, y \in [0, 1]$, where B_{2r} is the $2r$-th Bernoulli polynomial [173, 6]. If we let μ be the uniform measure on $[0, 1]$, the normalized eigenfunctions (of the integral operator) are 1, $c_m(\cdot) = \sqrt{2}\cos(2\pi m\,\cdot)$ and $s_m(\cdot) = \sqrt{2}\sin(2\pi m\,\cdot)$ for $m = 1, 2, \dots$, and the corresponding eigenvalues are 1 and m^{-2r} (both for c_m and s_m). Although the rectangle formula $f \mapsto n^{-1}\sum_{i=1}^n f(i/n)$ (a.k.a. `Uniform Grid` below) is known to be optimal for this kernel [183, 128] in the sense of worst-case error, this RKHS is commonly used for testing the efficiency of general kernel quadrature methods [6, 16, 83]. We also consider its multivariate extension on $[0, 1]^d$, i.e., the RKHS given by the product kernel $k_r^{\otimes d}(\boldsymbol{x}, \boldsymbol{y}) := \prod_{i=1}^d k_r(x_i, y_i)$ for $\boldsymbol{x} = (x_1, \dots, x_d), \boldsymbol{y} = (y_1, \dots, y_d) \in [0, 1]^d$.

We carried out the experiment for $(d, r) = (1, 1), (1, 3), (2, 1), (3, 3)$. For each (d, r), we compared the following algorithms for n-point quadrature rules with $n \in \{4, 8, 16, 32, 64, 128\}$.

N. + emp, N. + emp + opt: We used the functions $u_i^\top k(Z, \cdot)$ $(i = 1, \dots, n - 1)$ given by the Nyström approximation (4.10) with $s = n-1$ as test functions

[4] All done on a MacBook Pro, CPU: 2.4 GHz Quad-Core Intel Core i5, RAM: 8 GB 2133 MHz LPDDR3.



(a) $d = 1$, $r = 1$

(b) $d = 1$, $r = 3$

(c) $d = 2$, $r = 1$

(d) $d = 3$, $r = 3$

Figure 4.1: Periodic Sobolev spaces with kernel $k_r^{\otimes d}$: The average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 20 trials is plotted for each method of obtaining Q_n. The shaded regions show their standard deviation. The worst computational time per trial was 57 seconds of **Thin + opt** in $(d, r, n) = (3, 3, 128)$, where **Thinning** was 56 seconds and **N. + emp** [**+ opt**] was 22 seconds.

$\varphi_1, \dots, \varphi_{n-1}$ in Algorithm 4.1. The set Z was given as an $(\ell =)10n$-point independent sample from μ. We used $N = n^2$ samples from μ. In '**+ opt**' we additionally optimized the *convex* weights using (4.13)

M. + emp, M. + emp + opt ($d = 1$): We used the first $n-1$ functions of the sequence of eigenfunctions $1, c_1, s_1, c_2, s_2, \dots$ as test functions $\varphi_1, \dots, \varphi_{n-1}$ in Algorithm 4.1. We used $N = n^2$ samples from μ. In '**+ opt**' we additionally optimized the *convex* weights using (4.13).

Monte Carlo, iid Bayes: With an n-point independent sample $(x_i)_{i=1}^n$ from μ, we used uniform weights $1/n$ in **Monte Carlo** and the weights optimized using (4.13) in **iid Bayes**.

Uniform Grid ($d = 1$): We used the rectangle formula $f \mapsto n^{-1} \sum_{i=1}^n f(i/n)$. This is known to be optimal (not just up to constant, but exactly [183, 128]), and thus equivalent to the Bayesian quadrature on the uniform grid, i.e., the weights are already optimized.

Halton, Halton + opt ($d \geq 2$): For an n-point sequence given by the Halton sequence with Owen scrambling [64, 132], the uniform weights $w_i = 1/n$ is adopted in **Halton** and the weights are additionally optimized using (4.13) in **Halton + opt**.

Thinning, Thin + opt: Given an N-point independent sample $(y_i)_{i=1}^N$ with $N = n^2$ from μ, an n-point subset $(x_i)_{i=1}^n$ taken from a KT++ algorithm (kernel thinning [44, 45] combined with Compress++ algorithm [152] with the oversampling parameter $\mathfrak{g} = \min\{4, \log_2 n\}$, implemented with GoodPoints package: `https://github.com/microsoft/goodpoints`) is adopted in **Thinning**. In **'+ opt'** we additionally optimized the *convex* weights using (4.13).

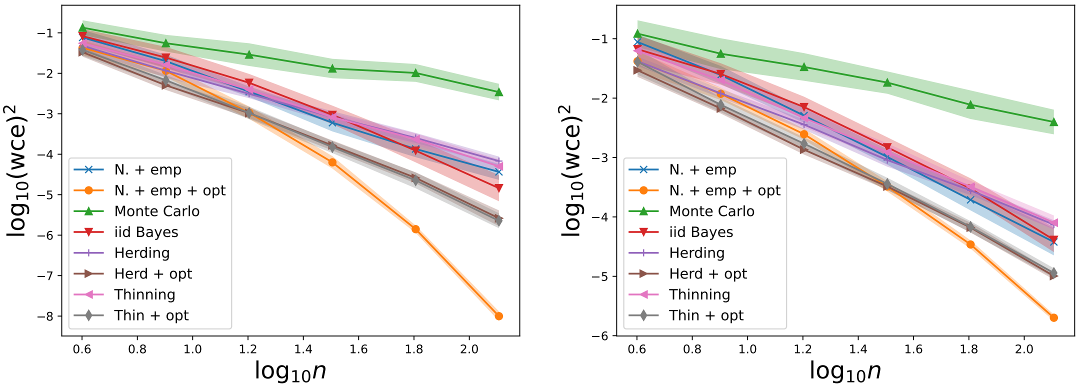
The results are given in Figure 4.1. In $d = 1$, the optimal rate given by **Uniform Grid** is known to be $\mathcal{O}(n^{-2r})$. As the uniform sampling is equal to the *optimized distribution* of Bach [6] in this case, **iid Bayes** also achieves this rate up to log factors. Although our theoretical guarantee for **M. + emp** is $\mathcal{O}(n^{1-2r} + N^{-1})$ with $N = n^2$ (Corollary 4.2), in the case $(d, r) = (1, 1)$, we can observe that in the experiment it is better than **iid Bayes** and close to the optimal error of **Uniform Grid**, but slightly worse than **Thinning**. Moreover, **N. + emp**, which does not use the information of spectral decomposition, is remarkably almost as accurate as **M. + emp** in $d = 1$. Furthermore, if we additionally use the knowledge of expectations, which **iid Bayes** is already doing, **M./N. + emp + opt** become surprisingly accurate even with $N = n^2$. They are worse than **Thinn + opt** when $r = 1$, but well outperform it when $r = 3$. Nonlinearity in the graph of these methods when $(d, r, n) = (1, 3, 128)$ should be from the numerical accuracy of the CQP solver (see also Section 4.E.1).

The accuracy of **N. + emp + opt** becomes more remarkable in multivariate cases. It behaves almost the same as **Halton + opt** in $d = 2$ and clearly beats it in $d = 3$. Also, the sudden jump of our methods around $n = 30$ in $(d, r) = (3, 3)$ seems to be caused by the jump in eigenvalues. Indeed, for the integral operator given by $k_3^{\otimes 3}$ with uniform measure, the eigenspace of the largest eigenvalue 1 is of dimension 27, and the next largest eigenvalue is 1/64. Again in the latter case, **N. + emp + opt** outperforms **Thin + opt**, and these results suggest that our method works better when there is a strong spectral decay, as is explicitly incorporated in our algorithm.

Note also that we can compare Figure 4.1 with Belhadji et al. [16, Figure 1] which includes some other methods such as DPPs, herding, and sequential Bayesian quadrature, as we did experiments under almost the same setting. In particular, in the case $(d, r) = (1, 3)$ where the eigenvalue decay is fast, we see that our method substantially outperforms the sequential Bayesian quadrature.

4.3.2 Measure reduction in machine learning datasets

We used two datasets from UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets/`). We set μ as the equally weighted measure over (a subset of) the data points $X = (X^{(1)}, \dots, X^{(d)})^\top$ ($d = 3, 5$, respectively), where each entry is centered and normalized. We considered the Gaussian kernel $\exp(-\|x - y\|^2/(2\lambda^2))$ whose hyperparameter λ is determined by *median heuristics* [50], and compared the performance of **N. + emp**, **N. + emp + opt** (with $\ell = 10n$, $N = n^2$), **Monte Carlo**, **iid Bayes**, **Thinning**, **Thin + opt**. We



(a) 3D Road Network data

(b) Power Plant data

Figure 4.2: Measure reduction in Gaussian RKHS with two ML datasets: The average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 20 trials is plotted for each method of obtaining Q_n. The shaded regions show their standard deviation. The worst computational time per trial was 14 seconds of **Thinning** [+ **opt**] in Power Plant data with $n = 128$, where **N. + emp** [+ **opt**] was 6.3 seconds.

also added **Herding**, an equally weighted greedy algorithm with global optimization [30], and its weight optimization **Herd + opt** within *convex* quadrature given by (4.13). Note that we chose the initial point for herding by uniform sampling, and this causes the randomness in Figure 4.5. We experimented $n \in \{4, 8, 16, 32, 64, 128\}$.

The first is **3D Road Network Data Set** [88]. The original dataset is 3-dimensional real vectors at 434874 points. To be able to compute the worst-case error (4.13) efficiently to evaluate each kernel quadrature, we used a random subset $\mathcal{X}$ of size $43487 = \lfloor 434874/10 \rfloor$ (fixed throughout the experiment) and defined μ as the uniform measure on it. We determined λ with the median heuristic by using a random subset of $\mathcal{X}$ with size 10000 and used the same $\mathcal{X}$ and λ throughout the experiment. The second is **Combined Cycle Power Plant Data Set** [89, 168]. The original dataset is 5-dimensional real vectors at 9568 points. We set the whole data as $\mathcal{X}$ and defined μ as the uniform measure on it. We determined λ with median heuristics by using the whole $\mathcal{X}$.

Figure 4.2 shows the results. We can observe that in both experiments **N. + emp + opt** successfully exploits the fast spectral decay of the Gaussian kernel and significantly outperforms other methods. Also, even without using the knowledge of any expectations, **N. + emp** (and **Thinning**) show a decent convergence rate comparable to **Herding** or **iid Bayes**, which actually use the additional information. See also the end of Section 4.E.2 for the plot of $\mathrm{wce}(Q_n; \mathcal{H}_k, \mu')$ for another set of empirical data μ'.

4.4 Concluding remarks

We leveraged a classical measure reduction tool, recombination, with spectral properties of kernels to construct kernel quadrature rules with positive weights. The resulting algorithms show strong performance within our experiments for small n (up to around 100) despite their restriction to convex weights. Our method has also recently been applied to Bayesian inference problems [1].

Although our method is applicable to fairly general situations, the usage or performance can be limited when it is difficult or inefficient to directly sample from the target measure μ. Hence, an interesting follow-up question, is how one could replace the i.i.d. samples with smarter sampling (DPP, importance sampling, etc) before the recombination is carried out. Further, our theoretical results do not fully explain the empirical superiority; especially the $1/\sqrt{\ell}$ term does not match the experiments and it is a challenging future research question to reduce this theoretical gap. Nevertheless, we believe that our Nyström-based method is the first generally applicable algorithm with convex weights with a guarantee from spectral decay, given access to i.i.d. samples from μ.

Appendix for Chapter 4

4.A Outline of Appendix

Appendix 4.B contains general results from which the results presented in the main text, in particular, Theorem 4.1, follow as special cases. Appendix 4.C contains the proofs of these theoretical results and needed technical lemmas. Appendix 4.D shows that if the expectations $\int \varphi_i(x) \,\mathrm{d}\mu(x)$ are known, then this knowledge can be used to further improve the theoretical bounds; it also gives a simple modification of Algorithm 4.1 doing this efficiently. Appendix 4.E provides additional numerical experiments and benchmarks.

4.B Theoretical results and remarks

In this section, we present theoretical results that include our main results as a special case. The proofs are given in Section 4.C.

Notation. For simplicity, for a quadrature Q_n given by points $(x_i)_{i=1}^n$ and weights $(w_i)_{i=1}^n$ and a probability measure μ, we denote the integration of an integrable function f on $\mathcal{X}$ with respect to these measures by

$$Q_n(f) = \sum_{i=1}^n w_i f(x_i), \qquad \mu(f) = \int_{\mathcal{X}} f(x) \, \mathrm{d}\mu(x),$$

respectively. We also write the inner product and norm of an RKHS $\mathcal{H}_k$ by $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ and $\|\cdot\|_{\mathcal{H}_k}$. Furthermore, we use the probability simplex Δ^n and convex hull $\operatorname{conv} A$ of a set $A \subset \mathbb{R}^d$ in the proofs:

$$\Delta^n := \left\{ (w_i)_{i=1}^n \,\middle|\, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}, \ \operatorname{conv} A := \left\{ \sum_{i=1}^n w_i a_i \,\middle|\, (w_i) \in \Delta^n, \, a_i \in A, \, n \geq 1 \right\}.$$

4.B.1 Quantitative results

We work under the following setting as in the assumption of Theorem 4.1.

Assumption A. *μ is a Borel probability measure on $\mathcal{X}$, and k is a positive definite kernel on $\mathcal{X}$ such that $\int_{\mathcal{X}} k(x,x) \, \mathrm{d}\mu(x) < \infty$. Further, k_0 is a positive definite kernel on $\mathcal{X}$ such that $k_1 := k - k_0$ is a positive definite kernel on $\mathcal{X}$.*

The following is a general result regarding a quadrature formula exactly integrating functions in $\mathcal{H}_{k_0}$.

Theorem 4.6. *Under Assumption A, if an n-point convex quadrature Q_n on $\mathcal{X}$ satisfies $Q_n(f) = \mu(f)$ for any $f = k_0(\cdot, x)$ with $x \in \mathcal{X}$, we have*

$$\operatorname{wce}(Q_n; \mathcal{H}_k, \mu) \leq Q_n(g) + \mu(g), \tag{4.14}$$

where g is the function given by $g(x) = \sqrt{k_1(x,x)}$. In particular, the following assertions hold for such a quadrature Q_n:

(a) We have $\mathrm{wce}(Q_n; \mathcal{H}_k, \mu) \le 2\|g\|_\infty = 2\sup_{x\in\mathcal{X}} \sqrt{k_1(x,x)}$.

(b) If we additionally have $Q_n(g) \le \mu(g)$*, then we have* $\mathrm{wce}(Q_n; \mathcal{H}_k, \mu) \le 2\mu(g)$.

(c) If we additionally have $Q_n(g^2) \le \mu(g^2)$ *instead of (b), we still have*

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \le 4 \int_{\mathcal{X}} k_1(x,x) \,\mathrm{d}\mu(x).$$

Remark 4.3. *For a Borel probability measure* ν *on* $\mathcal{X}$ *and a nonnegative function* $h : \mathcal{X} \to \mathbb{R}_{\ge 0}$*, we have an inequality* $\int_{\mathcal{X}} \sqrt{h(x)} \,\mathrm{d}\nu(x) \le \left(\int_{\mathcal{X}} h(x) \,\mathrm{d}\nu(x)\right)^{1/2}$*, so the above* $\mu(g)$ *can be upper bounded by* $\int_{\mathcal{X}} k_1(x,x) \,\mathrm{d}\mu(x)$*, which is equal to the trace of the integral operator given by* k_1*. Also, the assumption in Theorem 4.6 can be weakened to the integrability of* $\sqrt{k(x,x)}$ *from the same inequality as you can see in the proof.*

We can combine Theorem 4.6 with an empirical approximation of μ to obtain the following result, which essentially implies Theorem 4.1.

Theorem 4.7. *Under Assumption A, let* D_N *be a set of* N *independent samples from* μ*, and* $\widetilde{\mu}_N$ *be its empirical measure, i.e.,* $\widetilde{\mu}_N = \frac{1}{N}\sum_{y\in D_N} \delta_y$*. Then, if an* n*-point convex quadrature* Q_n *on* $\mathcal{X}$ *satisfies* $Q_n(f) = \widetilde{\mu}_N(f)$ *for any* $f = k_0(\cdot, x)$ *with* $x \in \mathcal{X}$*, we have*

$$\mathbb{E}\left[\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2\right] \le 2\mathbb{E}\left[(Q_n(g) + \widetilde{\mu}_N(g))^2\right] + \frac{2c_{k,\mu}}{N}, \tag{4.15}$$

where $g(x) \coloneqq \sqrt{k_1(x,x)}$ *and* $c_{k,\mu} \coloneqq \int_{\mathcal{X}} k(x,x) \,\mathrm{d}\mu(x) - \iint_{\mathcal{X}\times\mathcal{X}} k(x,y) \,\mathrm{d}\mu(x) \,\mathrm{d}\mu(y)$*. In particular, the following assertions hold for such a quadrature* Q_n*:*

(a) We have $\mathbb{E}[\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2] \le 8\sup_{x\in\mathcal{X}} k_1(x,x) + 2c_{k,\mu}/N$.

(b) If we additionally always require $Q_n(g) \le \widetilde{\mu}_N(g)$*, then we have*

$$\mathbb{E}\left[\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2\right] \le 8 \int_{\mathcal{X}} k_1(x,x) \,\mathrm{d}\mu(x) + \frac{2c_{k,\mu}}{N}.$$

The requirement $Q_n(g) \le \widetilde{\mu}_N(g)$ *can be replaced by* $Q_n(g^2) \le \widetilde{\mu}_N(g^2)$

Although we have assumed $k - k_0$ is positive definite in the previous assertions, the uniform bound works without the assumption as follows.

Proposition 4.8. *Let μ be a Borel probability measure on $\mathcal{X}$. Let k and k_0 be positive definite kernels on $\mathcal{X}$ satisfying $\int_{\mathcal{X}} \sqrt{k(x,x)}\,\mathrm{d}\mu(x), \int_{\mathcal{X}} \sqrt{k_0(x,x)}\,\mathrm{d}\mu(x) < \infty$. If an n-point convex quadrature Q_n on $\mathcal{X}$ satisfies $Q_n(f) = \mu(f)$ for any $f = k_0(\cdot, x)$ with $x \in \mathcal{X}$, we have*

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu) \leq 2 \sup_{x,y\in\mathcal{X}} \sqrt{|k(x,y) - k_0(x,y)|}.$$

Furthermore, if $\dim \mathcal{H}_{k_0} < n$, there exists an n-point convex quadrature Q_n satisfying $Q_n(f) = \mu(f)$ for each $f = k_0(\cdot, x)$.

In this chapter, we focus on the cases where k_0 is either given by the truncated Mercer decomposition or Nyström approximation. For many important kernels, however, we may also use the random Fourier features [139] or its periodic version [165] which can easily be combined with Proposition 4.8, but it is beyond the scope of this chapter to choose its appropriate variant for each kernel (see Liu et al. [105] for a list of variants).

4.B.2 Existence results

The existence of quadrature formulas satisfying the estimate of Theorem 4.6 or Theorem 4.7 is guaranteed when $\dim \mathcal{H}_{k_0} < n$.

Theorem 4.9. *Under Assumption A, if $\dim \mathcal{H}_{k_0} < n$, there exists an n-point convex quadrature Q_n satisfying $Q_n(f) = \mu(f)$ for each $f = k_0(\cdot, x)$. This still holds even if we additionally require $Q_n(g) \leq \mu(g)$ or $Q_n(g^2) \leq \mu(g^2)$ for $g(x) = \sqrt{k_1(x,x)}$.*

Remark 4.4. *This also implies the existence result of Q_n satisfying the condition in Theorem 4.7 if we replace μ by $\widetilde{\mu}_N$.*

The algorithm for constructing a kernel quadrature with Theorem 4.7 is given in the main body, see Algorithm 4.1. The ones with Theorem 4.6 requires further knowledge of the expectation of test functions, i.e., the values of $\int_{\mathcal{X}} \varphi_i(x)\,\mathrm{d}\mu(x)$ with $\mathcal{H}_{k_0} \subset \mathrm{span}\{\varphi_1, \dots, \varphi_{n-1}\}$. Under this additional information, we have an algorithm (Algorithm 4.2) based on random sampling given in the following section.

4.B.3 Eigenvalue estimate for Gaussian kernels

We provide proof of a folklore estimate on the eigenvalues of integral operators given by a Gaussian kernel. Let $k(x,y) = \exp(-\frac{1}{2\ell^2}(x-y)^2)$ for an $\ell > 0$ and $x, y \in \mathbb{R}$. Then, it has the following expansion [117, 86]:

$$k(x,y) = \sum_{m=0}^{\infty} \varphi_m(x)\varphi_m(y), \qquad \varphi_m(x) = \frac{1}{\ell^m\sqrt{m!}} x^m \exp\left(-\frac{x^2}{2\ell^2}\right). \tag{4.16}$$

Let μ be a Borel probability measure supported on a compact domain, i.e., $\mu(\{x \in \mathbb{R} \mid |x| > R\}) = 0$ for some $R > 0$. Let us consider the RKHS given by k over $\mathcal{X} := \operatorname{supp}\mu$.

Recall that σ_n is the n-th eigenvalue of the integral operator

$$\mathcal{K} : L^2(\mu) \to L^2(\mu); \quad f \mapsto \mathcal{K}f = \int_{\mathcal{X}} k(\cdot, y) f(y)\, \mathrm{d}\mu(y).$$

From the minimax property of eigenvalues of compact Hermitian operators,

$$\begin{aligned} \sigma_n &= \inf_{g_1,\ldots,g_{n-1}\in L^2(\mu)} \sup_{f \in L^2(\mu)\cap\{g_1,\ldots,g_{n-1}\}^\perp,\, \|f\|_{L^2(\mu)}=1} \langle f, \mathcal{K}f\rangle_{L^2(\mu)} \\ &\le \sup_{f\in L^2(\mu)\cap\{\varphi_0,\ldots,\varphi_{n-2}\}^\perp,\, \|f\|_{L^2(\mu)}=1} \langle f, \mathcal{K}f\rangle_{L^2(\mu)} \end{aligned}$$

hold, where the orthogonal complement is taken in terms of $L^2(\mu)$-inner product and φ_m are functions given in (4.16). They are indeed in $L^2(\mu)$ as μ is compactly supported. Now, let $k_n(x,y) := \sum_{m=n-1}^{\infty} \varphi_m(x)\varphi_m(y)$. For an $f \in L^2(\mu) \cap \{\varphi_0, \ldots, \varphi_{n-2}\}^\perp$, we have

$$\begin{aligned} \langle f, \mathcal{K}f\rangle_{L^2(\mu)} &= \iint_{\mathcal{X}\times\mathcal{X}} f(x)k(x,y)f(y)\,\mathrm{d}\mu(y)\,\mathrm{d}\mu(x) \\ &= \iint_{\mathcal{X}\times\mathcal{X}} f(x)k_n(x,y)f(y)\,\mathrm{d}\mu(y)\,\mathrm{d}\mu(x) \\ &\le \iint_{\mathcal{X}\times\mathcal{X}} f(x)\sqrt{k_n(x,x)}\sqrt{k_n(y,y)}f(y)\,\mathrm{d}\mu(y)\,\mathrm{d}\mu(x) \\ &\qquad\qquad \text{(positive definiteness)} \\ &= \left(\int_{\mathcal{X}} \sqrt{k_n(x,x)}f(x)\,\mathrm{d}\mu(x)\right)^2 \\ &\le \left(\int_{\mathcal{X}} k_n(x,x)\,\mathrm{d}\mu(x)\right) \|f\|_{L^2(\mu)}^2. \qquad \text{(Cauchy–Schwarz)} \end{aligned}$$

Therefore, we have the estimate $\sigma_n \leq \int_{\mathcal{X}} k_n(x,x)\,\mathrm{d}\mu(x)$. We have

$$k_n(x,x) = \sum_{m=n-1}^{\infty} \frac{1}{m!}\left(\frac{x}{\ell}\right)^{2m} \exp\left(-\left(\frac{x}{\ell}\right)^2\right),$$

and this can be regarded as the remainder term of the Maclaurin expansion, so there is a $\theta \in (0,1)$ such that

$$k_n(x,x) = \frac{1}{(n-1)!}\exp\left(-\theta\left(\frac{x}{\ell}\right)^2\right)\left(\frac{x}{\ell}\right)^{2(n-1)} \leq \frac{(x/\ell)^{2(n-1)}}{(n-1)!}.$$

In particular, if we have $|x| \leq R$ for μ-almost all x, we have a factorial decay $\sigma_n \leq \frac{(R/\ell)^{2(n-1)}}{(n-1)!}$.

4.B.4 Uniform robustness

In applications, the RKHS $\mathcal{H}_k$ may be misspecified and the quadrature rule μ^Q when computed for the misspecified function class $\mathcal{H}_k$ but applied to a function $f \notin \mathcal{H}_k$ leads only to the attainable bound

$$\begin{aligned}&\left|\int_{\mathcal{X}} f(x)\,\mathrm{d}\mu^Q(x) - \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x)\right| \\ &\leq \sup_{x\in\mathcal{X}}\left|f(x) - \widetilde{f}(x)\right|(|\mu^Q|_{\mathrm{TV}} + |\mu|_{\mathrm{TV}}) + \left\|\widetilde{f}\right\|_{\mathcal{H}_k} \mathrm{wce}(\mu^Q; \mathcal{H}_k, \mu)\end{aligned}$$

via triangle equality and standard integral estimates. Note that $|\cdot|_{\mathrm{TV}}$ denotes the total variation norm and the above applies to any $\widetilde{f} \in \mathcal{H}_k$; in particular, to the best approximation in uniform norm to f in $\mathcal{H}_k$. Since μ is a probability measure, $|\mu|_{\mathrm{TV}} = 1$ but if μ^Q is a signed measure with non-convex weights, its total variation $|\mu^Q|_{\mathrm{TV}}$ can be large, resulting in arbitrary large integration errors.

4.B.5 Equivalence between the projection/matrix Nyström approximations

Let k be a positive definite kernel on $\mathcal{X}$, $Z = (z_i)_{i=1}^{\ell} \subset \mathcal{X}$. Let $P_Z : \mathcal{H}_k \to \mathcal{H}_k$ be the projection operator onto $\mathrm{span}\{k(\cdot, z_i) \mid i = 1, \dots, \ell\}$. For arbitrary $x, y \in \mathcal{X}$, we can write

$$P_Z k(\cdot, x) = \sum_{i=1}^{\ell} a_i k(\cdot, z_i), \qquad P_Z k(\cdot, y) = \sum_{i=1}^{\ell} b_i k(\cdot, z_i),$$

where $a = (a_i)_{i=1}^\ell, b = (b_i)_{i=1}^\ell \in \mathbb{R}^\ell$. From the properties of projection, we have

$$k(z_j, x) = \langle k(\cdot, z_j), k(\cdot, x)\rangle_{\mathcal{H}_k} = \langle k(\cdot, z_j), P_Z k(\cdot, x)\rangle_{\mathcal{H}_k} = \sum_{i=1}^{\ell} a_i k(z_j, z_i).$$

In matrix notation, we have $k(Z, x) = k(Z, Z)a$, and $k(Z, y) = k(Z, Z)b$ from the same argument. Thus, by combining it with the property of Moore–Penrose inverse, we have

$$\begin{aligned} \langle P_Z k(\cdot, x), P_Z k(\cdot, y)\rangle_{\mathcal{H}_k} &= a^\top k(Z, Z)b \\ &= a^\top k(Z, Z)k(Z, Z)^+ k(Z, Z)b \qquad \text{(Moore–Penrose)} \\ &= k(x, Z)k(Z, Z)^+ k(Z, y). \end{aligned}$$

This shows the desired equivalence.

4.C Proofs

4.C.1 Proof of Theorem 4.6

Before proceeding to the proof of the theorem, we prepare a couple of assertions. The following is a well-known estimate proven by using the Cauchy–Schwarz inequality (see, e.g., Muandet et al. [123, Lemma 3.1 and its proof]).

Proposition 4.10. *Let k be a positive definite kernel on $\mathcal{X}$, and ν be a Borel probability measure with $\int_{\mathcal{X}} \sqrt{k(x, x)}\,\mathrm{d}\nu(x) < \infty$. Then, for each $f \in \mathcal{H}_k$, we have*

$$\left| \int_{\mathcal{X}} f(x)\,\mathrm{d}\nu(x) \right| \le \|f\|_{\mathcal{H}_k} \int_{\mathcal{X}} \sqrt{k(x, x)}\,\mathrm{d}\nu(x).$$

By using the proposition, we obtain the following technical lemma.

Lemma 4.11. *Let k and k_1 be a positive definite kernels on $\mathcal{X}$ such that $k - k_1$ is also positive definite. Let ν be a Borel probability measure on $\mathcal{X}$. Then, for any $n \ge 1$, $a_1, \dots, a_n \in \mathbb{R}$, $x_1, \dots, x_n \in \mathcal{X}$, if we let $f = \sum_{i=1}^n a_i k(\cdot, x_i)$ and $f_1 = \sum_{i=1}^n a_i k_1(\cdot, x_i)$, then we have*

$$\left| \int_{\mathcal{X}} f_1(x)\,\mathrm{d}\nu(x) \right| \le \|f\|_{\mathcal{H}_k} \int_{\mathcal{X}} \sqrt{k_1(x, x)}\,\mathrm{d}\nu(x).$$

Proof. From the positive definiteness of $k_0 \coloneqq k - k_1$, we have

$$\begin{aligned}\|f_1\|_{\mathcal{H}_{k_1}}^2 = \sum_{i,j=1}^n a_i a_j k_1(x_i, x_j) &\le \sum_{i,j=1}^n a_i a_j k_1(x_i, x_j) + \sum_{i,j=1}^n a_i a_j k_0(x_i, x_j) \\ &= \sum_{i,j=1}^n a_i a_j k(x_i, x_j) = \|f\|_{\mathcal{H}_k}^2.\end{aligned}$$

Hence it suffices to prove $|\nu(f_1)| \le \|f_1\|_{\mathcal{H}_{k_1}} \nu(g)$ for $g(x) \coloneqq \sqrt{k_1(x,x)}$, but it directly follows from Proposition 4.10. □

Proof of Theorem 4.6. Note first that, for each $f \in \mathcal{H}_{k_0}$, f is integrable with respect to μ. Indeed, we have

$$|f(x)| = |\langle f, k_0(\cdot, x)\rangle_{\mathcal{H}_{k_0}}| \le \|f\|_{\mathcal{H}_{k_0}} \|k_0(\cdot, x)\|_{\mathcal{H}_{k_0}} = \|f\|_{\mathcal{H}_{k_0}} \sqrt{k_0(x,x)} \le \|f\|_{\mathcal{H}_{k_0}} \sqrt{k(x,x)},$$

and it is integrable from assumption, so the equality $Q_n(f) = \mu(f)$ with $f = k_0(\cdot, x)$ is attained at a finite value.

Once we establish (4.14), item (b) is clear, and (a) follows from the fact that $Q_n(g)$ and $\mu(g)$ are both integrals of the function g with respect to a probability measure. Also, (c) is justified as follows:

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \le (Q_n(g)+\mu(g))^2 \le 2Q_n(g)^2+2\mu(g)^2 \le 2Q_n(g^2)+2\mu(g^2) \le 4\mu(g^2),$$

where $\mu(Q_n)^2 \le \mu(Q_n^2)$ and $\mu(g)^2 \le \mu(g^2)$ follows from the Cauchy–Schwarz.

To prove (4.14),we first prove

$$|Q_n(f) - \mu(f)| \le \|f\|_{\mathcal{H}_k} (Q_n(g) + \mu(g)) \tag{4.17}$$

for any f of the form $f = \sum_{i=1}^n a_i k(\cdot, x_i)$ with $n \ge 0$ and $a_1, \dots, a_n \in \mathbb{R}$. Given such an f, we have $Q_n(f_0) = \mu(f_0)$ for $f_0 \coloneqq \sum_{i=1}^n a_i k_0(\cdot, x_i)$ from the assumption. Thus, by letting $f_1 \coloneqq f - f_0 = \sum_{i=1}^n a_i k_1(\cdot, x_i)$, we have

$$Q_n(f) - \mu(f) = (Q_n(f) - \mu(f)) - (Q_n(f_0) - \mu(f_0)) = Q_n(f_1) - \mu(f_1).$$

As we have $|\nu(f_1)| \le \|f\|_{\mathcal{H}_k} \nu(g)$ for $\nu = Q_n, \mu$ from Lemma 4.11, we obtain $|Q_n(f_1) - \mu(f_1)| \le \|f\|_{\mathcal{H}_k}(Q_n(g) + \mu(g))$, and so (4.17) is shown for f of the form $f = \sum_{i=1}^n a_i k(\cdot, x_i)$.

Finally, we generalize (4.17) to any $f \in \mathcal{H}_k$. Let $\widetilde{f} \in \mathcal{H}_k$ can be written in the form $\sum_{i=1}^n a_i k(\cdot, x_i)$. If we let $h(x) = \sqrt{k(x,x)}$, from Proposition 4.10, we have

$$|Q_n(f - \widetilde{f})| \le \|f - \widetilde{f}\|_{\mathcal{H}_k} Q_n(h), \qquad |\mu(f - \widetilde{f})| \le \|f - \widetilde{f}\|_{\mathcal{H}_k} \mu(h).$$

Note that $\mu(h) < \infty$ follows from the integrability of $k(x,x)$ in Assumption A. Therefore, we have

$$\begin{aligned}
|Q_n(f) - \mu(f)| &\le |Q_n(\widetilde{f}) - \mu(\widetilde{f})| + |Q_n(f - \widetilde{f}) - \mu(f - \widetilde{f})| \\
&\le \|\widetilde{f}\|_{\mathcal{H}_k}(Q_n(g) + \mu(g)) + \|f - \widetilde{f}\|_{\mathcal{H}_k}(Q_n(h) + \mu(h)) \\
&\le \|f\|_{\mathcal{H}_k}(Q_n(g) + \mu(g)) + \|f - \widetilde{f}\|_{\mathcal{H}_k}(Q_n(g) + \mu(g) + Q_n(h) + \mu(h)).
\end{aligned}$$

As we can make $\|f - \widetilde{f}\|_{\mathcal{H}_k}$ arbitrarily small from the definition of $\mathcal{H}_k$, the proof of (4.14) is completed by taking the limit. □

4.C.2 Proof of Theorem 4.7

Proof. Denote $D_N = \{y_1, \ldots, y_N\}$ and note that the result follows from (4.15) and

$$\mathbb{E}\big[\widetilde{\mu}_N(g)^2\big] = \mathbb{E}\big[\widetilde{\mu}_N(g^2)\big] \le \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N g(y_i)^2\right] = \int_{\mathcal{X}} k_1(x,x)\,\mathrm{d}\mu(x), \tag{4.18}$$

where the first inequality is given by the Cauchy–Schwarz.

Indeed, (a) is an immediate consequence of (4.15) and Q_n and $\widetilde{\mu}_N$ making a probability measure, and (b) is obtained as $2\mathbb{E}[(Q_n(g) + \widetilde{\mu}_N(g))^2] \le 8\mathbb{E}[\widetilde{\mu}_N(g)^2] \le 8\int_{\mathcal{X}} k_1(x,x)\,\mathrm{d}\mu(x)$ by using (4.18) and the requirement $Q_n(g) \le \widetilde{\mu}_N(g)$.

When the requirement is $Q_n(g^2) \le \widetilde{\mu}_N(g^2)$, as we have $Q_n(g)^2 \le Q_n(g^2)$ and $\widetilde{\mu}_N(g)^2 \le \widetilde{\mu}_N(g^2)$ by the Cauchy–Schwarz, we also have by the AM–GM,

$$\begin{aligned}
2\mathbb{E}\big[(Q_n(g) + \widetilde{\mu}_N(g))^2\big] &\le 4\mathbb{E}\big[Q_n(g)^2\big] + 4\mathbb{E}\big[\widetilde{\mu}_N(g)^2\big] \\
&\le 4\mathbb{E}\big[Q_n(g^2)\big] + 4\mathbb{E}\big[\widetilde{\mu}_N(g^2)\big] \\
&\le 8\mathbb{E}\big[\widetilde{\mu}_N(g^2)\big] \le 8\int_{\mathcal{X}} k_1(x,x)\,\mathrm{d}\mu(x)
\end{aligned}$$

For showing (4.15), we remark that we always have

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \widetilde{\mu}_N) \le Q_n(g) + \widetilde{\mu}_N(g) \tag{4.19}$$

by applying Theorem 4.6 with $\widetilde{\mu}_N$ instead of μ.

Let $h(\mu), h(\widetilde{\mu}_N), h(Q_n) \in \mathcal{H}_k$ be the kernel mean embeddings of μ, $\widetilde{\mu}_N$ and μ^{Q_n}, i.e.,

$$h(\mu) := \int_{\mathcal{X}} k(\cdot, x)\,\mathrm{d}\mu(x), \quad h(\widetilde{\mu}_N) := \frac{1}{N}\sum_{i=1}^{N} k(\cdot, y_i), \quad h(Q_n) := \sum_{i=1}^{n} w_i k(\cdot, x_i),$$

where $(w_i)_{i=1}^n$ and $(x_i)_{i=1}^n$ are weights and points defining the quadrature Q_n. Remark that $h(\mu)$ is well-defined as $\int_{\mathcal{X}} k(x,x)\,\mathrm{d}\mu(x) < \infty$ [123, Lemma 3.1]. As we can rewrite the worst-case error as

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \widetilde{\mu}_N) = \|h(Q_n) - h(\widetilde{\mu}_N)\|_{\mathcal{H}_k}, \qquad \mathrm{wce}(Q_n; \mathcal{H}_k, \mu) = \|h(Q_n) - h(\mu)\|_{\mathcal{H}_k},$$

by triangle inequality and the AM–GM, we obtain

$$\begin{aligned}\mathbb{E}\big[\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2\big] &\le \mathbb{E}\big[(\mathrm{wce}(Q_n; \mathcal{H}_k, \widetilde{\mu}_N) + \|h(\mu) - h(\widetilde{\mu}_N)\|_{\mathcal{H}_k})^2\big] \\ &\le 2\mathbb{E}\big[\mathrm{wce}(Q_n; \mathcal{H}_k, \widetilde{\mu}_N)^2\big] + 2\mathbb{E}\big[\|h(\mu) - h(\widetilde{\mu}_N)\|_{\mathcal{H}_k}^2\big] \\ &\le 2\mathbb{E}\big[(Q_n(g) + \widetilde{\mu}_N(g))^2\big] + 2\mathbb{E}\big[\|h(\mu) - h(\widetilde{\mu}_N)\|_{\mathcal{H}_k}^2\big],\end{aligned}$$

where we have used (4.19) in the last inequality. It now suffices to prove $\mathbb{E}\big[\|h(\mu) - h(\widetilde{\mu}_N)\|_{\mathcal{H}_k}^2\big] = c_{k,\mu}/N$ for showing (4.15).

Indeed, we have

$$\begin{aligned}&\mathbb{E}\big[\|h(\mu) - h(\widetilde{\mu}_N)\|_{\mathcal{H}_k}^2\big] = \mathbb{E}\big[\|h(\mu)\|_{\mathcal{H}_k}^2\big] - 2\mathbb{E}\big[\langle h(\mu), h(\widetilde{\mu}_N)\rangle_{\mathcal{H}_k}\big] + \mathbb{E}\big[\|h(\mu)\|_{\mathcal{H}_k}^2\big] \\ &= \iint_{\mathcal{X}\times\mathcal{X}} k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) - \frac{2}{N}\sum_{i=1}^{N}\int_{\mathcal{X}} \mathbb{E}[k(x,y_i)]\,\mathrm{d}\mu(x) + \frac{1}{N^2}\sum_{i,j=1}^{N}\mathbb{E}[k(y_i,y_j)] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}[k(y_i,y_i)] + \left(1 - 2 + \frac{N(N-1)}{N^2}\right)\iint_{\mathcal{X}\times\mathcal{X}} k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) = \frac{c_k}{N},\end{aligned}$$

since $\iint_{\mathcal{X}\times\mathcal{X}} k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) = \int_{\mathcal{X}} \mathbb{E}[k(x,y_i)]\,\mathrm{d}\mu(x) = \mathbb{E}[k(y_i,y_j)]$ holds for $i \neq j$. Thus, the proof is completed. □

4.C.3 Proof of Proposition 4.8

Proof. As Q_n exactly integrates the functions in $\mathcal{H}_{k_0}$, we have $\mathrm{wce}(Q_n; \mathcal{H}_{k_0}, \mu) = 0$. So, if we set $Q_n(f) = \sum_{i=1}^n w_i f(x_i)$, then we have, from (4.13) with kernel k_0,

$$0 = \sum_{i,j=1}^n w_i w_j k_0(x_i, x_j) - 2\sum_{i=1}^n w_i \int_{\mathcal{X}} k_0(x_i, y)\,\mathrm{d}\mu(y) + \iint_{\mathcal{X}\times\mathcal{X}} k_0(x, y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y). \tag{4.20}$$

If we extract this from the formula (4.13) for the kernel k, we have, by letting $k_1 \coloneqq k - k_0$,

$$\begin{aligned}\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 &= \mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 - \mathrm{wce}(Q_n; \mathcal{H}_{k_0}, \mu)^2 \\ &= \sum_{i,j=1}^n w_i w_j k_1(x_i, x_j) - 2\sum_{i=1}^n w_i \int_{\mathcal{X}} k_1(x_i, y)\,\mathrm{d}\mu(y) \\ &\quad + \iint_{\mathcal{X}\times\mathcal{X}} k_1(x, y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y).\end{aligned}$$

So, if we define $M \coloneqq \sup_{x\in\mathcal{X}}|k_1(x, y)| = \sup_{x\in\mathcal{X}}|k(x, y) - k_0(x, y)|$, we have

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \le \left(\sum_{i,j=1}^n w_i w_j M + 2\sum_{i=1}^n w_i \int_{\mathcal{X}} M\,\mathrm{d}\mu(y) + \iint_{\mathcal{X}\times\mathcal{X}} M\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\right) = 4M,$$

as Q_n is a convex quadrature. The existence follows from almost the same proof as in the proof of Theorem 4.6, but in this case, it directly follows from Tchakaloff's thorem [162, 12]. □

4.C.4 Proof of Theorem 4.9

Proof. We prove the existence of the version $Q_n(g) \le \mu(g)$. The other follows just by replacing every g in the proof below by g^2.

Let $\varphi_1, \ldots, \varphi_{n-1} \in \mathcal{H}_{k_0}$ satisfy $\mathcal{H}_{k_0} = \mathrm{span}\{\varphi_1, \ldots, \varphi_{n-1}\}$. Also, let $y, y_1, y_2, \ldots$ be independent samples from μ. Now, consider the vector-valued function $\boldsymbol{\psi} = (\varphi_1, \ldots, \varphi_{n-1}, g)^\top \in \mathbb{R}^n$. Note that $\mathbb{E}[\|\boldsymbol{\psi}(y)\|] < \infty$ follows from the integrability of elements in $\mathcal{H}_{k_0}$ and g with respect to μ. Therefore, by [66, Theorem 11], with probability 1, there exists an N such that $\mathbb{E}[\boldsymbol{\psi}(y)] \in \mathrm{conv}\{\boldsymbol{\psi}(y_1), \ldots, \boldsymbol{\psi}(y_N)\}$. So, in particular, there exist deterministic points $x_1, \ldots, x_N \in \mathcal{X}$ satisfying $\mathbb{E}[\boldsymbol{\psi}(y)] \in$

$\mathrm{conv}\{\boldsymbol{\psi}(x_1),\dots,\boldsymbol{\psi}(x_N)\}$. For such $(x_i)_{i=1}^N$, consider an optimal solution that is also a *basic* feasible solution of the following linear programming problem:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^N w_i g(x_i) \\ &\text{subject to} && [\boldsymbol{\varphi}(x_1)\cdots\boldsymbol{\varphi}(x_N)]\,\boldsymbol{w} = \int_{\mathcal{X}} \boldsymbol{\varphi}(x)\,\mathrm{d}\mu(x),\ \ \boldsymbol{w}\ge \mathbf{0}, \end{aligned} \tag{4.21}$$

where $\boldsymbol{\varphi} = (1, \varphi_1, \dots, \varphi_{n-1})^\top \in \mathbb{R}^n$ is another vector-valued function (note that its first coordinate is constant so that any feasible solution of (4.21) sums up to one). Such a basic solution $\boldsymbol{w}$ has at most n nonzero entries, say $(w_{i_1},\dots,w_{i_n}) \in \Delta^n$ with $1 \le i_1 < \cdots < i_n \le N$. Then, the quadrature Q_n given by weights $(w_{i_j})_{j=1}^n$ and points $(x_{i_j})_{j=1}^n$ satisfies $Q_n(\boldsymbol{\varphi}) = \mu(\boldsymbol{\varphi})$ and $Q_n(g) \le \mu(g)$. The latter follows from the optimality of $\boldsymbol{w}$ and the fact that $\mathbb{E}[\psi(y)] \in \mathrm{conv}\{\psi(y_1),\dots,\psi(y_N)\}$ leads to a feasible solution with the objective $\mathbb{E}[g(y)] = \mu(g)$). □

4.C.5 Proof of Theorem 4.3

We prove the theorem by using an existing bound regarding the Nystöm approximation for matrices, which is more common in the machine learning literature.

Let $A = (A_{ij})_{i,j=1}^N \in \mathbb{R}^{N\times N}$ be a symmetric positive semi-definite matrix. Let us denote it as $A = [\boldsymbol{a}_1,\dots,\boldsymbol{a}_N]$ by using $\boldsymbol{a}_1,\dots,\boldsymbol{a}_N \in \mathbb{R}^N$. Then, we independently sample $i_1,\dots,i_\ell$ from $\{1,\dots,N\}$ uniformly, and construct a submatrix $B = (a_{i_j i_k})_{j,k=1}^\ell$. If we let B_s be the best rank-s approximation of B and B_s^+ its pseudoinverse, the matrix

$$\widetilde{A} = [\boldsymbol{a}_{i_1},\dots,\boldsymbol{a}_{i_\ell}] B_s^+ \begin{bmatrix} \boldsymbol{a}_{i_1} \\ \vdots \\ \boldsymbol{a}_{i_\ell} \end{bmatrix} \tag{4.22}$$

works as a rank-s approximation of A.

We use the following result on this matrix version:

Proposition 4.12 ([97, Theorem 2]). *For a positive semi-definite matrix A, the rank-s approximation $\widetilde{A}$ given above satisfies, with probability at least $1-\delta$, the following:*

$$\|A - \widetilde{A}\|_2 \le \|A - A_s\|_2 + \frac{2N}{\sqrt{\ell}} A_{\max}\left(1 + \sqrt{\frac{D^A_{\max}}{A_{\max}} \frac{N-\ell}{N-1/2} \frac{1}{\beta(\ell,N)} \log\frac{1}{\delta}}\right),$$

where $\beta(\ell, N) = 1 - \frac{1}{2\max\{\ell, N-\ell\}}$, $A_{\max} = \max_i A_{ii}$, $D^A_{\max} = \max_{i,j}(A_{ii} + A_{jj} - 2A_{ij})$ *and* A_s *is the best rank-s approximation of* A.

As $D^A_{\max} \le 2A_{\max}$, if we have $N \ge 2\ell$, it holds that

$$\frac{D^A_{\max}}{A_{\max}} \frac{N-\ell}{N-1/2} \frac{1}{\beta(\ell, N)} \le 2 \frac{N-\ell}{N-1/2} \frac{N-\ell-1/2}{N-\ell} \le 2,$$

and we can just state

$$\|A - \widetilde{A}\|_2 \le \|A - A_s\|_2 + \frac{2N}{\sqrt{\ell}} A_{\max} \left(1 + \sqrt{2\log\frac{1}{\delta}}\right). \tag{4.23}$$

We show the following lemma as a consequence of this proposition.

Lemma 4.13. *Let* $s \le \ell$ *be positive integers and* $\delta > 0$. *Let* $k : \mathcal{X} \times \mathcal{X}$ *be a symmetric and positive definite kernel and* $y_1, y_2, \ldots$ *be i.i.d. random variables taking values in* $\mathcal{X}$. *For each* N, *define the* $N \times N$ *matrices* $K(N), K_s(N), K^Z_s(N)$ *by*

$$K(N)_{ij} = \frac{k(y_i, y_j)}{N}, \quad K_s(N)_{ij} = \frac{1}{N}\sum_{m=1}^{s} \sigma_m e_m(y_i) e_m(y_j), \quad K^Z_s(N)_{ij} = \frac{k^Z_s(y_i, y_j)}{N},$$

where $Z = (y_1, \ldots, y_\ell)$.

Then, there exists a sequence $\varepsilon_N \to 0$ *such that*

$$\|K(N) - K^Z_s(N)\|_2 \le \|K(N) - K_s(N)\|_2 + \frac{2\sup_x k(x,x)}{\sqrt{\ell}} \left(1 + \sqrt{2\log\frac{1}{\delta}}\right) \tag{4.24}$$

is met with probability at least $1 - \delta - \varepsilon_N$.

Proof. We assume $N \ge 2\ell$. Let $i_1, \ldots, i_\ell$ be independent uniform samples from $\{1, \ldots, N\}$. Consider the event E_N that $i_1, \ldots, i_\ell$ are all different. Then, $\mathbb{P}(E_N) = \prod_{i=1}^{\ell} \frac{N+1-i}{N}$ converges to 1 as $N \to \infty$, and let $\varepsilon_N = 1 - \mathbb{P}(E_N)$. By using Proposition 4.12, (4.24) and $\max_i K(N)_{ii} \le N^{-1}\sup_x k(x,x)$, we have that the probability

$$\mathbb{P}\left(\|K(N) - \widetilde{K}_s(N)\|_2 \le \|K(N) - K_s(N)\|_2 + \frac{2\sup_x k(x,x)}{\sqrt{\ell}} \left(1 + \sqrt{2\log\frac{1}{\delta}}\right) \,\middle|\, E_N \right)$$

is at least $(1-\delta-\varepsilon_N)/\mathbb{P}(E_N) \geq 1-\delta-\varepsilon_N$, where $\widetilde{K}_s(N)$ is the rank-s Nyström approximation of the matrix $K(N)$ by using indices $i_1,\ldots,i_\ell$. From (4.22), if we take $\widetilde{W} = k(y_{i_j}, y_{i_k})_{j,k=1}^{\ell}$ and $\widetilde{W}_s$ its best rank-s approximation, it actually satisfies

$$\widetilde{K}_s(N)_{ij} = \frac{1}{N} k(y_i, D)\widetilde{W}_s^+ k(D, y_j) = \frac{1}{N} k_s^D(y_i, y_j),$$

where $D = (y_{i_1}, \ldots, y_{i_\ell})$ and k_s^D is the Nyström approximation given in the main body.

As $y_1, \ldots, y_N$ are i.i.d. samples, we can see that $(Z, (y_i)_{i=1}^N)$ (without any conditioning) and $(D, (y_i)_{i=1}^N)$ conditioned on E_N actually have the same distribution, so we are done. □

We finally prove the result for the Nyström approximation of integral operators.

Proof of Theorem 4.3. Take a sufficiently large N and let us use $K(N), K_s(N), K_s^Z(N)$ defined in the previous lemma with $y_1, y_2, \ldots$ independently sampled from μ.

It suffices to consider the case $C_k := \sup_{x\in\mathcal{X}} k(x,x) < \infty$. It is clear that $K_s(N)_{ii} \leq K(N)_{ii} \leq C_k/N$, and from (4.9), we also have

$$k_s^Z(x,x) = k(x,Z)W_s^+ k(Z,x) \leq k(x,Z)W^+ k(Z,x) = \|P_Z k(\cdot,x)\|_{\mathcal{H}_k}^2 \leq \|k(\cdot x)\|_{\mathcal{H}_k}^2 = k(x,x),$$

and so $K_s^Z(N)_{ii} \leq C_k/N$.

For a matrix $A(N) \in \mathbb{R}^{N\times N}$ defined by $A(N)_{ij} = (1-\delta_{ij})(K(N) - K_s^Z(N))$, i.e., the matrix given by deleting the diagonal, we have $\|A(N)\|_2 \to \|\mathcal{K}_s^Z - \mathcal{K}\|$ as $N \to \infty$ almost surely [93, Theorem 3.1]. Since we have observed that $\|K(N) - K_s^Z(N) - A(N)\|_2 \leq C_k/N$, we have

$$\|K(N) - K_s^Z(N)\|_2 \to \|\mathcal{K}_s^Z - \mathcal{K}\|, \qquad N \to \infty$$

almost surely. The same argument yields $\|K(N) - K_s(N)\|_2 \to \sigma_{s+1}$, as it converges to the norm of the integral operator given by the kernel $\sum_{m\geq s+1} \sigma_m e_m(x) e_m(y)$.

Now, by letting A_N be the event that (4.24) holds (so $\mathbb{P}(A_N) \geq 1-\delta-\varepsilon_N$), the desired inequality (4.11) almost surely holds under the event $\limsup A_N = \bigcap_{N>\ell}\bigcup_{M\geq N} A_M$. Indeed, under this event, we can just take the limit of both sides of (4.24) for an appropriate subsequence of $(2\ell, 2\ell+1, \ldots)$. As we have

$$\mathbb{P}(\limsup A_N) = \lim_{N\to\infty} \mathbb{P}\left(\bigcup_{M\geq N} A_N\right) \geq \lim_{N\to\infty}(1-\delta-\varepsilon_N) = 1-\delta,$$

the proof is completed. □

4.D Kernel quadrature when expectations are known

When we use an approximate kernel k_0 and know the exact expectations of test functions $\varphi_1, \dots, \varphi_{n-1}$ with $\mathcal{H}_{k_0} \subset \operatorname{span}\{\varphi_1, \dots, \varphi_{n-1}\}$, we can obtain an n-point kernel quadrature that exactly integrates $\varphi_1, \dots, \varphi_{n-1}$ by Algorithm 4.2.

Algorithm 4.2 Kernel quadrature with random convex hulls

```
Input: A positive definite kernel $k$ on $\mathcal{X}$, a probability measure $\mu$ on $\mathcal{X}$, integers
    $N \geq n \geq 1$, another kernel $k_0$ and functions $\varphi_1, \dots, \varphi_{n-1}$ on $\mathcal{X}$ with $\mathcal{H}_{k_0} \subset$
    $\operatorname{span}\{\varphi_1, \dots, \varphi_{n-1}\}$
Output: With some probability, returns $Q_n \coloneqq \{(w_i, x_i) \mid i = 1, \dots, n\} \subset \mathbb{R} \times \mathcal{X}$
    with $(w_i) \in \Delta^n$
1: Calculate the expectations $\int_{\mathcal{X}} \varphi_1(x)\,\mathrm{d}\mu(x), \dots, \int_{\mathcal{X}} \varphi_{n-1}(x)\,\mathrm{d}\mu(x)$
2: Sample $y_1, \dots, y_N$ independently from $\mu$
3: For a vector-valued function $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_{n-1})^\top$ and $k_{1,\mathrm{diag}}(x) = k(x,x) -$
   $k_0(x,x)$, solve the linear programming problem ($|\cdot|_0$ denotes the number of
   nonzero entries)
```

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{w}^\top k_{1,\mathrm{diag}}(\boldsymbol{x}) \\ \text{subject to} \quad & [\boldsymbol{\varphi}(y_1) \cdots \boldsymbol{\varphi}(y_N)]\, \boldsymbol{w} = \int_{\mathcal{X}} \boldsymbol{\varphi}(x)\,\mathrm{d}\mu(x), \\ & \boldsymbol{w} \geq \mathbf{0},\ \mathbf{1}^\top \boldsymbol{w} = 1,\ |\boldsymbol{w}|_0 \leq n. \end{aligned} \tag{4.25}$$

```
   to obtain points $\{x_1, \dots, x_n\} \subset \{y_1, \dots, y_N\}$ and weights $(w_i) \in \Delta^n$ satisfying
```

$$\sum_{i=1}^n w_i \boldsymbol{\varphi}(x_i) = \int_{\mathcal{X}} \boldsymbol{\varphi}(x)\,\mathrm{d}\mu(x)$$

```
   if (4.25) is feasible.
```

We make several remarks on this algorithm. First, the problem (4.25) is, strictly speaking, not a linear programming (LP) problem, as it includes the sparsity constraint $|\boldsymbol{w}|_0 \leq n$. However, as it only contains n equality constraints, its *basic feasible solution* always satisfies $|\boldsymbol{w}|_0 \leq n$ and the simplex algorithm automatically gives such a sparse (and optimal) solution even if we do not explicitly impose this constraint, so we call it an LP for simplicity. Second, *this algorithm occasionally fails to output Q_n* as, with some probability, the LP has no feasible solution.

Although we can repeat the algorithm until we succeed, the number N should be chosen appropriately. See Remark 4.5 for this point. Finally, our algorithm has possibly related approaches such as sparse optimization and Sard's method, see Remark 4.6 and 4.7.

Remark 4.5. *A simple approach for constructing a quadrature formula [66] was recently proposed: randomly sample candidate points and find a solution by using a linear programming (LP) solver. Indeed, for an independent sample $y_1, \ldots, y_N \sim \mu$, we can construct a quadrature formula with convex weights exactly integrating the functions in $\mathcal{F} = \operatorname{span}\{\varphi_1, \ldots, \varphi_{n-1}\}$ using a subset of these points if and only if we have*

$$\int_{\mathcal{X}} \boldsymbol{\varphi}(x) \, \mathrm{d}\mu(x) \in \operatorname{conv}\{\boldsymbol{\varphi}(y_1), \ldots, \boldsymbol{\varphi}(y_N)\}, \tag{4.26}$$

where $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_{n-1})^\top : \mathcal{X} \to \mathbb{R}^{n-1}$ and $\operatorname{conv} A$ denotes the convex hull of A. Several sharp estimates for the probability of the event (4.26) *are available in Chapter 2 & 3. Under the event* (4.26)*, we can find a desired rule by using the simplex method for the LP problem* (4.25)*.*

Remark 4.6. *From the viewpoint of subsampling, a direct way to obtain quadrature formulas with convex weights supported on a small number of points, is to first sample N candidate points $D_N = (x_1, \ldots, x_N)$ and then solve the following sparse optimization problem:*

$$\begin{array}{ll} \textit{minimize} & \boldsymbol{w}^\top k(D_N, D_N)\boldsymbol{w} - 2\boldsymbol{w}^\top \int_{\mathcal{X}} k(D_N, y) \, \mathrm{d}\mu(y) \\ \textit{subject to} & \boldsymbol{w} \geq \mathbf{0}, \ \mathbf{1}^\top \boldsymbol{w} = 1, \ |\boldsymbol{w}|_0 \leq n, \end{array} \tag{4.27}$$

where $k(D_N, D_N)$ is the corresponding $N \times N$ Gram matrix. Unfortunately, exactly solving this problem is computationally challenging, in particular in contrast to our approach that exploits the spectral properties of k and μ. Nevertheless, one could use sparse optimization to obtain an approximate solution of (4.27)*: although the simplex constraint ($\boldsymbol{w} \geq \mathbf{0}$, $\mathbf{1}^\top \boldsymbol{w} = 1$) makes it impossible to exploit the classical ℓ_1 regulatization, there are possible alternatives under this constraint [137, 98, 102] or use the DC (difference of convex functions) algorithm to incorporate the sparsity constraint to find local minima [57]. This is a promising research direction, and our general sample estimates might provide a first step toward this direction.*

Remark 4.7. *Sard's method [147, 99] for constructing numerical integration rules uses the n degree of freedom (of choosing weights in our setting) separately; m ($\leq n$) for exactness over a certain m-dimensional space of test functions, and the remaining $n - m$ for minimizing an error criterion such as the worst-case error. In the context of kernel quadrature, one way to use Sard's method with exactness over $\mathcal{F}$ (an m-dimensional space of test functions) is as follows [85, 156]:*

$$\begin{array}{ll} \textit{minimize} & \boldsymbol{w}^\top k(D_n, \boldsymbol{x}_n)\boldsymbol{w} - 2\boldsymbol{w}^\top \int_{\mathcal{X}} k(\boldsymbol{x}_n, y)\,\mathrm{d}\mu(y) \\ \textit{subject to} & \boldsymbol{w}^\top f(D_n) = \int_{\mathcal{X}} f(y)\,\mathrm{d}\mu(y), \ \forall f \in \mathcal{F}, \end{array} \tag{4.28}$$

where $D_n = (x_1, \ldots, x_n)^\top$, $f(D_n) = (f(x_1), \ldots, f(x_n))^\top$. This amounts to solving a convex quadratic programming for $\boldsymbol{w}$ in an $(n-m)$-dimensional subspace of $\mathbb{R}^n$ (without constraint). This is similar to our approach in that it enforces exactness in a certain finite-dimensional space of test functions. One key difference is that Sard's approach aims for a quadrature formula on a given set of points, whereas our method determines also the points themselves. Hence, the combination of these two approaches seems to be an interesting future research topic.[5]

Computational complexity. A tricky part of this approach, essentially based on random convex hulls, is that the algorithm possibly does not output a quadrature formula. Hence, the following quantity plays an important role to estimate the essential complexity of the algorithm:

$$N_\varphi = \inf \left\{ N \geq 1 \,\middle|\, \mathbb{P}(\mathbb{E}[\boldsymbol{\varphi}(y)] \in \operatorname{conv}\{\boldsymbol{\varphi}(y_1), \ldots, \boldsymbol{\varphi}(y_N)\}) \geq \frac{1}{2} \right\},$$

where $y, y_1, y_2, \ldots$ are independent samples from μ. This value is known to be finite and estimated under a variety of conditions on $\boldsymbol{\varphi}(y)$ (see Chapter 2 & 3 for details). If we have some knowledge of μ, we can just keep trying the algorithm with $N = N_\varphi$ until it succeeds, and its expected computational time is $\mathcal{O}(nN_\varphi + C(n, N_\varphi))$, where $C(a, b)$ is the (expected) cost of solving an $a \times b$ LP with a simplex method. Note that, though the worst-case computational time of the simplex method is

[5] For example, we can pick the first m eigenfunctions of the integral operator as test functions, and find n points and weights that minimize the worst-case error while exactly integrating the test functions from a larger set of candidate points. An obvious challenge is that quadratic programming does not supply sparsity, whereas the approach of this chapter has been fully based on the sparsity of a basic feasible solution of an LP problem.

exponential, it is empirically $\mathcal{O}(ab\min\{a,b\})$ in practice [133, 151]. In addition, $N_\varphi = \mathcal{O}(n)$ holds in examples with some symmetry [178, 66], so in that case we have a heuristic complexity estimate of $\mathcal{O}(n^3)$.

Choice of approximate kernels. Similarly to the empirical version discussed in the main text, we prove quantitative estimates when k_0 is given by the Mercer approximation or Nyström approximation. Remark that the necessary information for using these methods is different. Whereas using the Mercer approximation requires the knowledge of Mercer decomposition $k(x,y) = \sum_{m=1}^\infty \sigma_m e_m(x)e_m(y)$ and their exact integration $\int_\mathcal{X} e_m(x)\,\mathrm{d}\mu(x)$, the Nyström approximation only requires the exact integral values of the kernel, $\int_\mathcal{X} k(x,y)\,\mathrm{d}\mu(y)$, and so is more generally applicable. See the following sections for details.

In the following, we assume that the kernel attains the Mercer decomposition $k(x,y) = \sum_{m=1}^\infty \sigma_m e_m(x)e_m(y)$, where $\sigma_1 \ge \sigma_2 \ge \cdots \ge 0$ and $(e_m)_{m=1}^\infty$ is an orthonormal set of $L^2(\mu)$.

4.D.1 Algorithm 4.2 with Mercer approximation

If we use the truncated Mercer decomposition as an approximate kernel, we have the following result.

Theorem 4.14. *If Algorithm 4.2 with $k_0 := \sum_{m=1}^{n-1} \sigma_m e_m(x)e_m(y)$ and $\varphi_i = e_i$ successfully outputs a convex quadrature Q_n, then it satisfies the following:*

(a) If $C := \sup_{m\ge1}\|e_m\|_\infty < \infty$, we have $\mathrm{wce}(Q_n;\mathcal{H}_k,\mu)^2 \le 4C^2\sum_{m=n}^\infty \sigma_m$.

(b) As N in Algorithm 4.2 tends to infinity, we have

$$\mathbb{P}\left(\mathrm{wce}(Q_n;\mathcal{H}_k,\mu)^2 \le 4\sum_{m=n}^\infty \sigma_m\right) \to 1.$$

Proof. As in the proof of Corollary 4.2, $k - k_0$ is positive definite. Thus, when $\sum_{m=n}^\infty \sigma_m < \infty$, the kernel and the measure μ satisfies Assumption A. So Theorem 4.6(a) implies (a) of this theorem, since $k_1(x,x) = \sum_{m=n}^\infty \sigma_m e_m(x)^2 \le C^2\sum_{m=n}^\infty \sigma_m$.

For (b), if we have $Q_n(k_{1,\mathrm{diag}}) \le \mu(k_{1,\mathrm{diag}})$, then Theorem 4.6 implies

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2 \le 4 \int_{\mathcal{X}} k_1(x, x)\, \mathrm{d}\mu(x) = \int_{\mathcal{X}} \sum_{m=n}^{\infty} \sigma_m e_m(x)^2\, \mathrm{d}\mu(x) = 4 \sum_{m=n}^{\infty} \sigma_m.$$

So it suffices to prove $\mathbb{P}(Q_n(k_{1,\mathrm{diag}}) \le \mu(k_{1,\mathrm{diag}})) \to 1$ as $N \to \infty$, and it is shown by considering the optimal basic feasible solution of the LP (4.25) and the following fact [66, Theorem 11]:

$$\mathbb{P}(\mathbb{E}[\boldsymbol{\psi}(y)] \in \mathrm{conv}\{\boldsymbol{\psi}(y_1), \dots, \boldsymbol{\psi}(y_N)\}) \to 1, \qquad N \to \infty,$$

where $\psi = (\varphi_1, \dots, \varphi_{n-1}, k_{1,\mathrm{diag}})^\top$. Indeed, if $\boldsymbol{\psi}(y) \in \mathrm{conv}\{\boldsymbol{\psi}(y_1), \dots, \boldsymbol{\psi}(y_N)\}$, the LP becomes feasible and $Q_n(k_{1,\mathrm{diag}}) \le \mu(k_{1,\mathrm{diag}})$ follows from optimality. See the proof of Theorem 4.9 (Section 4.C.4) for a more detailed explanation. □

Note that the boundedness of C is a typical assumption [see 109, Assumption 3.2 and references therein], while it does not necessarily hold [118, Section 3]. Under some assumptions, we can quantify the probability that the LP (4.25) becomes feasible.

Sampling bound. Suppose 1 is an eigenfunction of $\mathcal{K}$. This is satisfied, e.g., in the following cases:

- μ is a Haar measure on a compact group and k is shift-invariant.
- k is a kernel based on Stein's identity [129, 156, 5] with respect to μ.

In this case, we have a theoretical bound of the required N in Algorithm 4.2 as follows.

Theorem 4.15. *Suppose* 1 *is an eigenfunction of* $\mathcal{K}$, *i.e.,* $\int_{\mathcal{X}} k(\cdot, y)\, \mathrm{d}\mu(y)$ *is a constant function. Then, for each* $n \ge 2$ *and* $N \ge 6(n-1) \sup_{x\in\mathcal{X}} \sum_{m=1}^{n-1} e_m(x)^2$, *Algorithm 4.2 returns a feasible quadrature with probability at least* $1 - 2^{1-n}$, *i.e., for an independent sample* $y_1, \dots, y_N$ *from* μ, *we have*

$$\mathbb{P}\left(\int_{\mathcal{X}} \boldsymbol{\varphi}(x)\, \mathrm{d}\mu(x) \in \mathrm{conv}\{\boldsymbol{\varphi}(y_1), \dots, \boldsymbol{\varphi}(y_N)\} \right) \ge 1 - \frac{1}{2^{n-1}},$$

where $\boldsymbol{\varphi} = (e_1, \dots, e_{n-1})^\top$. *If the value* $C = \sup_{m\ge1} \|e_m\|_\infty$ *is finite,* $N \ge 6C(n-1)^2$ *is also sufficient for the above estimate.*

Proof. This follows from Theorem 2.12 and Theorem 2.15. □

4.D.2 Algorithm 4.2 with Nyström approximation

Although the method discussed in the previous section requires the exact information on Mercer decomposition, if we make use of the Nyström approximation, we only require the values of $\int_{\mathcal{X}} k(x, y) \,\mathrm{d}\mu(y)$ for $x \in \mathcal{X}$.

Recall that $k_s^Z(x, y)$ is the rank-s Nyström approximation of the kernel k based on the point set $Z = (z_1, \dots, z_\ell)$. From (4.10), we can use $\varphi_i^Z \coloneqq u_i^\top k(Z, \cdot)$ as test functions.

Theorem 4.16. *Let $n \le \ell$ and $\delta > 0$, and let Z be an ℓ-point independent sample from μ. If Algorithm 4.2 with $k_0 = k_{n-1}^Z$ and $\varphi_i = \varphi_i^Z$ successfully outputs a convex quadrature Q_n, then with probability at least $1 - \delta$, we have*

$$\operatorname{wce}(Q_n; \mathcal{H}_k, \mu)^2 \le 4n\sigma_n + 4 \sum_{m=n+1}^{\infty} \sigma_m + \frac{8(n-1)\sup_{x\in\mathcal{X}} k(x, x)}{\sqrt{\ell}} \left(1 + \sqrt{2 \log \frac{1}{\delta}}\right)$$

Proof. As in the proof of Corollary 4.4, $k - k_{n-1}^Z$ is positive definite. Also, we can assume $\sum_{m=1}^\infty \sigma_m < \infty$, as otherwise the right-hand side is infinity. Thus k and $k_0 = k_{n-1}^Z$ satisfy Assumption A.

Note that for a function of the form $c(x, y) = a \cdot b(x)b(y)$ with $a \in \mathbb{R}$ and $b \in L^2(\mu)$, and an orthonormal set $(f_i)_{i\in I} \subset L^2(\mu)$ of $L^2(\mu)$ with $b \in \overline{\operatorname{span}\{f_i \mid i \in I\}}$, we have

$$\begin{aligned}\sum_{i\in I} \iint_{\mathcal{X}\times\mathcal{X}} f_i(x)c(x, y)f_i(y) \,\mathrm{d}\mu(x) \,\mathrm{d}\mu(y) &= a \sum_{i\in I} \langle b, f_i\rangle_{L^2}^2 \\ &= a\|b\|_{L^2}^2 = \int_{\mathcal{X}} c(x, x) \,\mathrm{d}\mu(x). \qquad (4.29)\end{aligned}$$

If we here use the orthonormal set $(e_m)_{m=1}^\infty$ that appears in the Mercer decomposition, by letting $k_1 \coloneqq k - k_{n-1}^Z$ and using the linear extension of (4.29), we have

$$\begin{aligned}\int_{\mathcal{X}} k_1(x,x)\,\mathrm{d}\mu(x) &= \int_{\mathcal{X}} (k(x,x) - k_0(x,x))\,\mathrm{d}\mu(x)\\ &= \sum_{m=1}^{\infty} \left\langle e_m, (\mathcal{K} - \mathcal{K}_{n-1}^Z)e_m \right\rangle_{L^2}\\ &\le \sum_{m=n}^{\infty} \langle e_m, \mathcal{K}e_m\rangle_{L^2} + \sum_{m=1}^{n-1} \|\mathcal{K} - \mathcal{K}_{n-1}^Z\| \|e_m\|_{L^2}^2\\ &= \sum_{m=n}^{\infty} \sigma_m + (n-1)\|\mathcal{K} - \mathcal{K}_{n-1}^Z\|. \qquad (4.30)\end{aligned}$$

Then, by combining this with Theorem 4.3 and Theorem 4.6(c), we obtain the desired estimate. □

Remark 4.8. *If we denote by N_φ the required number of samples, the computational complexity of the above algorithm becomes $\mathcal{O}(n\ell N_\varphi + n\ell^2 + C(n, N_\varphi))$, including the cost of computing the Nyström approximation as well as test functions at N_φ samples (see also Remark 4.2).*

4.E Additional numerical experiments

In this section, we provide additional experiments on Algorithm 4.2 using random convex hulls, as well as the approximated version of the **N. + emp** described in Remark 4.2. Section 4.E.1 shows the comparison of Algorithm 4.2 (with Mercer/Nyström approximation) with some of the methods mentioned in the main text under the periodic Sobolev spaces with uniform measure. Section 4.E.2 investigates Algorithm 4.2 (with Nystöm approximation) as well as the approximate but fast algorithm for **N. + emp**, under the setting of empirical measure reduction.

4.E.1 Periodic Sobolev spaces with uniform measure

We conducted experiments under the same setting as in Section 4.3.1, except that we additionally have the following methods:

Nyström, Nyström + opt: We used the same test functions as **N. + emp** with the random set Z of size $\ell = 10n$, but for Algorithm 4.2. We used

$N = 10n$ samples for the LP (4.25). In **Nyström + opt** we additionally optimized the convex weights using (4.13).

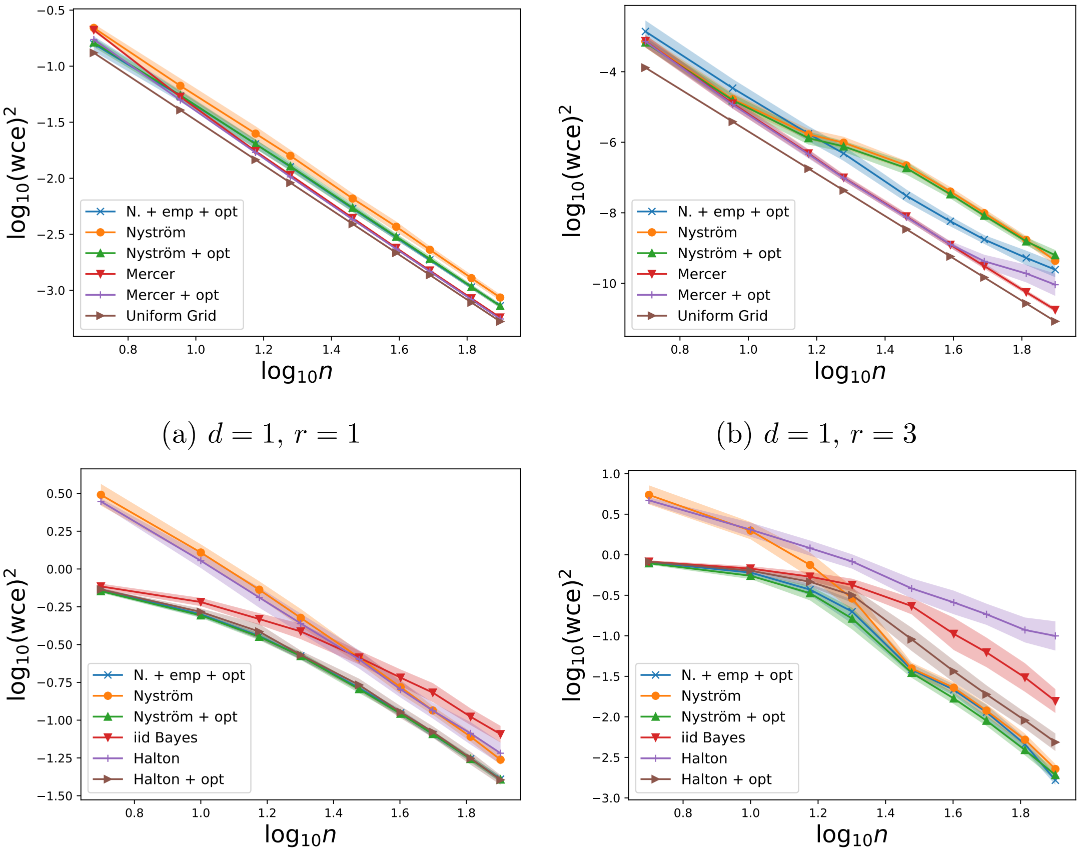
Mercer, Mercer + opt ($d = 1$): We used the same test functions as **M. + emp**, but for Algorithm 4.2. We used $N = 10n$ samples for the LP (4.25). In **Mercer + opt** we additionally optimized the convex weights using (4.13).

The results are given in Figure 4.3. The weights of **Nyström** and **Mercer** are already almost optimized as they exactly integrate a certain family of functions, so the additional CQP (4.13) does not change the error so much. Surprising is that **N. + emp + opt** is almost as good as **Nyström + opt** or even better. This implies that the recombination points with respect to a moderately large ($N = n^2$ in this case) empirical measure can provide a good convergence rate in Bayesian quadrature [78], even though the (equally weighted) empirical measure itself is not that close to the true measure.

Odd behavior of 'Mercer'. As we can see in Figure 4.3(a,b), the methods based on the exact Mercer decomposition becomes very close to optimal when $n = 15, 65$. As it seemed to be caused by the parity of n, we carried out another experiment for $n \in \{5, 9, 15, 19, 29, 39, 49, 65, 79\}$ (Figure 4.4), then **Mercer** and its optimization clearly became the best methods except the exact optimal **Unifrom Grid**. It might be related to the structure of the periodic Sobolev space, that has, for each eigenvalue except for 1, two-dimensional eigenspace (cos and sin), but needs further investigation. Also, in the case $(d, r) = (1, 3)$, we see '**+ opt**' make the quadrature less accurate for a big n, but it is theoretically almost impossible, so it seems to be caused by the numerical accuracy of the CQP solver.

4.E.2 Measure reduction in machine learning datasets

We conducted experiments under the same setting as in Section 4.3.2. We additionally adopted **Nyström, Nyström + opt** (with $N = 20n$), and **FNE, FNE + opt**, where **FNE** (stands for 'fast N. + emp') is the approximate algorithm for **N. + emp** by omitting the inequality in (4.5) and using the randomized SVD [63] (see Remark 4.2).



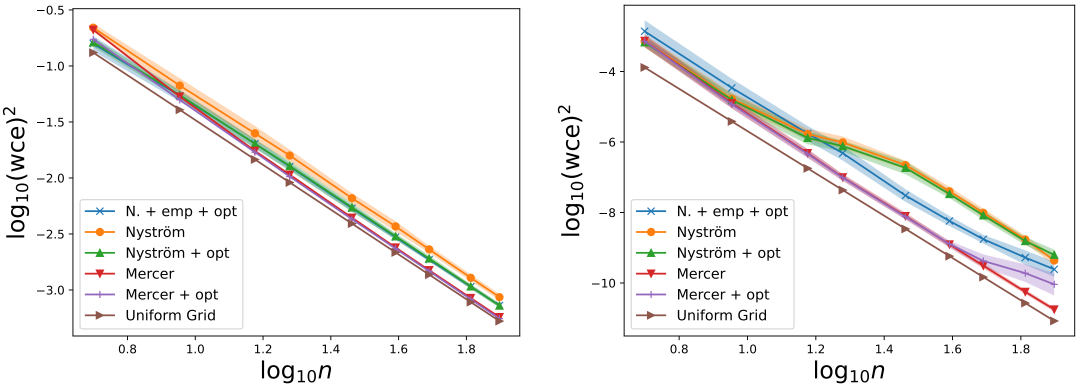
(a) $d = 1$, $r = 1$

(b) $d = 1$, $r = 3$

(c) $d = 2$, $r = 1$

(d) $d = 3$, $r = 3$

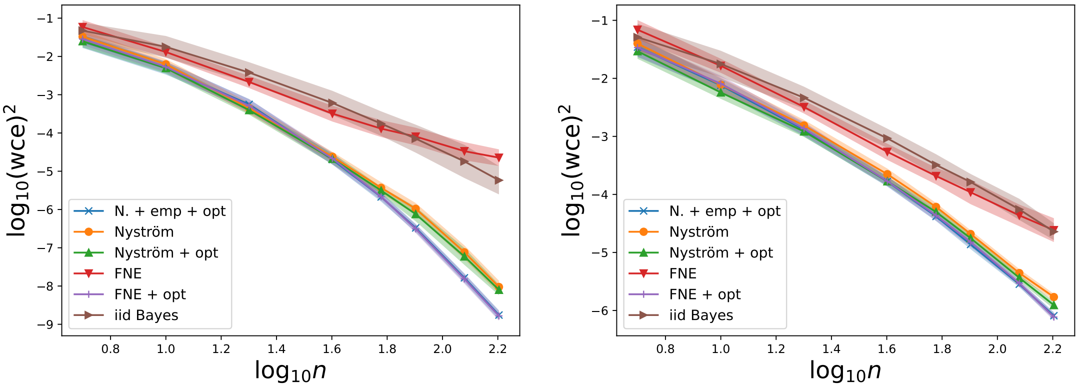
Figure 4.3: Periodic Sobolev spaces with kernel $k_r^{\otimes d}$: The average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 50 trials is plotted for each method of obtaining Q_n. The shaded regions are sample standard deviations. The worst computational time per trial was 5 seconds of **N. + emp** and **N. + emp + opt** in $(d, r, n) = (3, 3, 80)$, while **Nyström** and **Nyström + opt** ran in 0.9 seconds under the same setting. There were no infeasible LPs.



(a) $d = 1$, $r = 1$

(b) $d = 1$, $r = 3$

Figure 4.4: Supplemental experiments for Figure 4.3. n is all odd.



(a) 3D Road Network data

(b) Power Plant data

Figure 4.5: Measure reduction in Gaussian RKHS with two ML datasets: The average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 50 trials is plotted for each method of obtaining Q_n. The shaded regions are sample standard deviations. The worst computational time per trial was 13 seconds of **N. + emp** and **N. + emp + opt** in 3D Road Network data with $n = 160$. There were 7 infeasible LPs (and 800 feasible LPs) in the experiment (a) with **Nyström** or **Nyström + opt**. There were no infeasible LPs in (b).

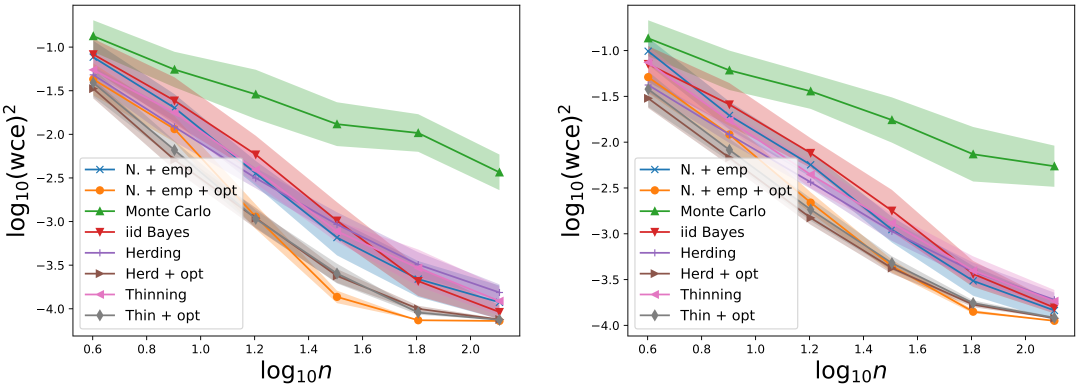
The results are given in Figure 4.5. **N. + emp + opt** and **FNE + opt** show almost the same convergence. While in the largest case $n = 160$, the average runtime of (**N. + emp + opt**, **FNE + opt**) was $(13.0, 2.07)$ seconds in 3D Road Network data and $(12.8, 2.09)$ seconds in Power Plant data, respectively. Although our theoretical guarantee no longer holds for **FNE**, it accelerates the algorithm while surprisingly maintaining the accuracy. **Nystöm** or **Nyström + opt** behave much better than **iid Bayes**, but are slightly less accurate than **N. + emp + opt** and **FNE + opt**, whereas they have good theoretical guarantees (Theorem 4.16). Their computational time was basically between that of **FNE + opt** and **N. + emp + opt**.

Comparison with another empirical measure. The setting of 'ML datasets' treated here is empirical measures given by some real data, so it is also just an approximation of a true distribution from the viewpoint of frequentists. Therefore, if we want to evaluate the performance of measure reduction methods with regard to the true distribution, we should measure the worst-case error using it. As it is not feasible in reality, we take another empirical measure μ' (of the same size as but different from the empirical measure μ, used in the construction of a kernel quadrature rule Q_n), and plot the quantities of $\mathrm{wce}(Q_n; \mathcal{H}_k, \mu')$ to better estimate the actual performance of Q_n in this section.

The overall setting is the same as in Section 4.3.2, except for the following points:

- In the 3D Road Network Data Set, we used another random 43487-point subset from the remaining $434874 - 43487$ data points to define μ'.
- In the Combined Cycle Power Plant Data Set, we used exactly half of the whole data points to define μ (so the size of $\operatorname{supp} \mu$ is different from the original experiment) and the other half to define μ'.

Note that μ and μ' were randomly taken at first and fixed throughout the experiment. The median heuristics as well as the normalization of the data (for both of the points in μ and μ') was carried out by using the statistical information solely given by μ.



(a) 3D Road Network data

(b) Power Plant data

Figure 4.6: Measure reduction in Gaussian RKHS with two ML datasets with another empirical measure: The average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu')^2)$ over 20 trials is plotted for each method of obtaining Q_n. The shaded regions show their standard deviation. The worst computational time per trial was 14 seconds of **Thinning** [**+ opt**] in Power Plant data with $n = 128$, where **N. + emp** [**+ opt**] was 6.2 [6.1] seconds.

The results are given in Figure 4.6. We can see that, though our methods are still competitive, the error eventually becomes dominated by the (MMD-)distance between μ and μ' as n gets larger. This is inevitable as we are only using the empirical measure μ to construct Q_n, so in an application to this kind of setting, we can just pick any method whose error is sufficiently small compared to the 'inevitable' error.

Chapter 5

Sampling-based Nyström approximation and kernel quadrature

We analyze the Nyström approximation of a positive definite kernel associated with a probability measure. We first prove an improved error bound for the conventional Nyström approximation with i.i.d. sampling and singular-value decomposition in the continuous regime; the proof techniques are borrowed from statistical learning theory. We further introduce a refined selection of subspaces in Nyström approximation with theoretical guarantees that is applicable to non-i.i.d. landmark points. Finally, we discuss their application to convex kernel quadrature and give novel theoretical guarantees as well as numerical observations.

5.1 Introduction

Kernel methods form a prominent part of modern machine-learning tools. However, making kernel methods scalable to large datasets is an ongoing challenge. The main bottleneck is that the kernel Gram matrix scales quadratically in the number of data points. For large-scale problems, the number of matrix entries can easily be of the order of hundreds-thousands or millions so even storing the full Gram matrix can become too costly. Several approaches have been developed to deal with these, among the most prominent are the Random Fourier Features and the Nyström method. In this chapter, we revisit and generalize the Nyström

Table 5.1: Main quantitative results. Individual bounds are available in Remark 5.1, Theorem 5.11, and Proposition 5.13. For the explanation on each kernel, see at the end of **Contribution** section. Here are remarks on the notation. (a) σ_i is the i-th eigenvalue of the integral operator $\mathcal{K} : L^2(\mu) \to L^2(\mu); g \mapsto \int_{\mathcal{X}} k(\cdot, x)g(x)\,\mathrm{d}\mu(x)$. (b) μ_X denotes the equally weighted empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ given by $X = (x_i)_{i=1}^N$. (c) $\mu(\cdot)$ and $\mu_X(\cdot)$ denote the integrals over the diagonal. See (5.4).

Quantity	Bound	Assumption
$\mathbb{E}[\mu(\sqrt{k - k_s^Z})]$ $\mathbb{E}[\mu(k - k_s^Z)]$	$\mathcal{O}\left(\sqrt{\sum_{i>s}\sigma_i} + \frac{(\log \ell)^{2d+1}}{\ell}\right)$	$\begin{cases} Z \sim_{\mathrm{iid}} \mu,\ k\text{: bounded} \\ \sigma_i \lesssim \exp(-\beta i^{1/d}) \end{cases}$
$\mu(\sqrt{k^Z - k_{s,\mu}^Z})^2$ $\mu(k^Z - k_{s,\mu}^Z)$	$\sum_{i>s}\sigma_i$	Z: fixed
$\mathbb{E}[\mu_X(\sqrt{k^Z - k_{s,X}^Z})]^2$ $\mathbb{E}[\mu_X(k^Z - k_{s,X}^Z)]$	$\sum_{i>s}\sigma_i$	Z: fixed, $X \sim_{\mathrm{iid}} \mu$

method and provide new error estimates. Consequences are theoretical guarantees for kernel quadrature and improvements on the standard Nyström method that goes beyond uniform subsampling of data points.

Nyström approximation. While already discussed in Section 4.2.3, the main idea of the Nyström method is to replace the original positive definite kernel k with another kernel k_{app} that is constructed by random projection of the elements in the (in general infinite-dimensional) RKHS associated with k into a low-dimensional RKHS. A consequence of this is that the Gram matrix of k_{app} is a low-rank approximation of the original Gram matrix. Concretely, let μ denote a probability measure on a (Hausdorff) space $\mathcal{X}$ and k a kernel on $\mathcal{X}$; then the standard Nyström approximation uses the random kernel

$$k^Z(x, y) := k(x, Z)k(Z, Z)^+k(Z, y). \tag{5.1}$$

where $Z = (z_i)_{i=1}^{\ell}$ is an ℓ-point subset of $\mathcal{X}$ usually taken i.i.d. from μ [43, 97].

Further s-rank approximation. While less common, the following rank-reduced version is of interest in the context of kernel quadrature, as explained in Section 4.2.3:

$$k_{\mathrm{app}}(x, y) = k_s^Z(x, y) := k(x, Z)k(Z, Z)_s^+ k(Z, y), \tag{5.2}$$

where $k(Z, Z)_s^+$ is the Moore–Penrose pseudo-inverse of the best s-rank approximation of the Gram matrix $k(Z, Z) = (k(z_i, z_j))_{i,j=1}^{\ell}$ with $s \le \ell$. Note that $k_\ell^Z = k^Z$.

Let us briefly review our motivation for this rank reduction coming from kernel quadrature. If we are given an s-rank kernel k_{app} and a probability measure μ, by Tchakaloff's theorem there is a discrete probability measure ν supported over at most $s+1$ points satisfying $\int_{\mathcal{X}} f \,\mathrm{d}\mu = \int_{\mathcal{X}} f \,\mathrm{d}\nu$ for all $f \in \mathcal{H}_{k_{\mathrm{app}}}$, where $\mathcal{H}_{k_{\mathrm{app}}}$ is the finite-dimensional RKHS associated with the kernel k_{app}. Such a measure ν works as a kernel quadrature rule if the k_{app} well approximates the original kernel k, and the rank s directly affects the number of (possibly expensive) function evaluations we need to estimate each integral. The primary error criterion in this chapter is

$$\int_{\mathcal{X}} \sqrt{k(x, x) - k_{\mathrm{app}}(x, x)} \,\mathrm{d}\mu(x), \tag{5.3}$$

which arises from the error estimate in kernel/Bayesian quadrature (Chapter 4 and [1]).

Contribution. Our first theoretical result is that the expectation of (5.3) is of the order $\mathcal{O}\big(\sqrt{\sum_{i>s}\sigma_i} + \mathrm{polylog}(\ell)/\ell\big)$ when the eigenvalues $(\sigma_i)_{i=1}^{\infty}$ of the kernel integral operator induced by (k, μ) enjoy exponential convergence (the expectation is taken over the empirical sample Z). Key to the proof of this bound is the use of concepts from statistical learning theory; in particular, the (local) Rademacher complexity. This error estimate is far better than the bound $\mathcal{O}\big(\text{spectral term} + s^{1/2}/\ell^{1/4}\big)$ that follows from the existing high-probability estimate $\int_{\mathcal{X}}(k(x, x) - k_s^Z(x, x)) \,\mathrm{d}\mu(x) = \mathcal{O}(s\sigma_s + \sum_{i>s}\sigma_i + s/\sqrt{\ell})$ (see the proof of Corollary 4.4). Combining our new bound with known kernel quadrature estimates explains the strong empirical performance of the convex kernel quadrature given in Chapter 4; the theoretical bounds in Chapter 4 were not even better than Monte-Carlo in terms of ℓ.

Our second contribution is the use of other k_{app} than k_s^Z with better bounds of (5.3), for a general class of landmark points Z rather than just an i.i.d. sample from μ. This generalization allows to use other sets Z in (5.2) to achieve better overall performance; e.g. sampling Z from determinantal point processes (DPPs) on $\mathcal{X}$ is known to be advantageous in applications. To construct and provide theoretical guarantees for such improved Nyström constructions we revisit and generalize a method that was proposed in Santin and Schaback [146] and give further theoretical guarantees applicable to kernel quadrature rules.

The following is the list of low-rank approximations presented in the chapter:

- k^Z and k_s^Z: Usual Nyström approximations using landmark points Z. See (5.1) and (5.2).
- $k_{s,\mu}^Z$: The s-rank truncated Mercer decomposition of the kernel k^Z with respect to the measure μ. See (5.11).
- $k_{s,X}^Z$: A version of $k_{s,\mu}^Z$ with μ given by the empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ of the set $X = (x_i)_{i=1}^N$. This actually coincides with k_s^Z when $X = Z$; see (5.6).

See Table 5.1 for a summary of our quantitative results.

Outline. Section 5.2 discusses the existing literature and introduces some notation. Section 5.3 contains our first main result, namely the analysis of k_s^Z for an i.i.d. Z; Appendix 5.A provides the necessary background from statistical learning theory. In Section 5.4, we then treat a general Z to give refined low-rank approximations together with theoretical guarantees, rather than the conventional k_s^Z. In Section 5.5, we discuss how our bounds yields new theories and methods for the recent random kernel quadrature construction, which enables us to explain the empirical performance as well as to build some strong candidates whose performance is assessed by numerical experiments. All the omitted proofs are given in Appendix 5.B.

5.2 Related literature and notation

To simplify the notation, we denote

$$\nu(f) := \int_{\mathcal{X}} f(x)\,\mathrm{d}\nu(x), \quad \nu(h) := \int_{\mathcal{X}} h(x,x)\,\mathrm{d}\nu(x) \tag{5.4}$$

for any functions $f : \mathcal{X} \to \mathbb{R}$, $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a (probability) measure ν on $\mathcal{X}$, whenever the integrals are well-defined. In this notation, the aim of this chapter is to bound $\mu(\sqrt{k - k_{\mathrm{app}}})$ or $\mu(k - k_{\mathrm{app}})$ for a class of low-rank approximation k_{app}. Also, A^+ denotes the Moore–Penrose pseudo-inverse of a matrix A.

Approximation of the Gram matrix. The standard use of the Nyström method in ML is to replace the Gram matrix $k(X,X)$ for a set $X = (x_i)_{i=1}^N$ by the low-rank matrix $k^Z(X,X)$ where k^Z is defined as in (5.1). A well-developed literature studies the case when $Z = (z_i)_{i=1}^\ell$ is uniformly and independently sampled from X, see Drineas et al. [43], Kumar et al. [97], Yang et al. [181], Jin et al. [80], Li et al. [101]. Further, the cases of leverage-based sampling [54], DPPs [100], and kernel K-means samples [130] have received attention. Moreover, two variants of the standard Nyström method have been studied: the first replaces the Moore-Penrose inverse of $k(Z,Z)$ in (5.1) with the pseudo-inverse of the best s-rank approximation of $k(Z,Z)$ as in (5.2) via SVD [43, 97, 101]; the second uses the best s-rank approximation of $k^Z(X,X)$, see [166, 177]. For a brief overview in this regard, see Wang et al. [177, Remark 1].

Approximation of the integral operator. The matrix $k(X,X)$ can be regarded as a finite-dimensional representation of the linear (integral) operator

$$\mathcal{K} : L^2(\mu) \to L^2(\mu), \quad (\mathcal{K}f)(x) = \int_{\mathcal{X}} k(x,y) f(y)\,\mathrm{d}\mu(y).$$

We denote with $(\sigma_i, e_i)_{i=1}^\infty$ the eigenpairs of the operator $\mathcal{K}$, and assume the eigenvalues are ordered $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. The Mercer decomposition exists under mild assumptions (for example, $\operatorname{supp}\mu = \mathcal{X}$, k is continuous and $\int_{\mathcal{X}} k(x,x)\,\mathrm{d}\mu(x) < \infty$ [160] are sufficient) and gives the representation

$$k(x,y) = \sum_{i=1}^\infty \sigma_i e_i(x) e_i(y), \tag{5.5}$$

where $\|e_i\|_{L^2(\mu)} = 1$, and $(\sqrt{\sigma_i} e_i)_{i=1}^{\infty}$ is an orthonormal basis of the RKHS $\mathcal{H}_k$ of k. Hence, a natural approach is to just truncate this expansion after s terms, $k_{\text{app}} = \sum_{i=1}^{s} \sigma_i e_i(x) e_i(y)$, to get a finite-dimensional approximation of the kernel k. This approach is natural since the approximation quality of the operator $\mathcal{K}$ determines the resulting error estimates. Unfortunately, it is often rendered useless since the Mercer decomposition depends on the tuple (k, μ), and while explicit expression is known for special choices, in general, it is unlikely to have a closed-form representation of the eigenpairs $(\sigma_i, e_i)_{i=1}^{\infty}$.

Other approximations. A compromise that is relevant to our work is proposed in Santin and Schaback [146]. Instead of using the Mercer decomposition of $\mathcal{K}$ one uses the Mercer decomposition of (5.1). Our main result allows to generalize this approach and to provide theoretical guarantees missing in the reference. Related is the paper Gauthier [51] that studies the interactions of several Hilbert-Schmidt spaces of (integral) operators given by a Nyström approximation/projection of a kernel-measure pair as in the present chapter; further, Chatalic et al. [29] considers a low-rank approximation of an empirical kernel mean embedding by using a Nyström-based projection. The leverage-based sampling studied in Gittens and Mahoney [54] has continuous counterparts. One with a slight modification is in the kernel literature [6], while the exact counterpart can be found in a context from approximation theory [31] under the name of *optimally-weighted sampling*, which essentially proposes sampling from $s^{-1} \sum_{i=1}^{s} e_i^2(x) \, \mathrm{d}\mu(x)$.

The power function. Finally, the square root of the diagonal term $\sqrt{k(x,x) - k^Z(x,x)}$ or its generalization is known as the *power function* in the literature on kernel-based interpolation [38, 145, 87]. There the primary interest is its L^∞ (uniform) norm, rather than the $L^1(\mu)$ norm, $\mu(\sqrt{k - k_{\text{app}}})$, or the $L^2(\mu)$ norm, $\mu(k - k_{\text{app}})$, that appear in kernel quadrature estimates and error estimates of the Nyström/Mercer type decompositions.

Kernel quadrature. The literature on kernel quadrature other than Chapter 4 includes herding [30, 7, 78, 167], weighted/correlated sampling [6, 16, 17, 15, 46], a subsampling method called thinning [44, 45, 152]. We refer to Table 4.1 in the previous chapter for a comparison of existing algorithms in terms of their convergence guarantees and computational complexities.

5.3 Analyzing k_s^Z for i.i.d. Z via statistical learning theory

Let $Z = (z_i)_{i=1}^{\ell} \subset \mathcal{X}$ and k_s^Z be the s-dimensional kernel given by $k_s^Z(x,y) = k(x,Z)k(Z,Z)_s^+k(Z,y)$ as in the usual Nyström approximation. Throughout the chapter, suppose we are provided the singular value decomposition of the matrix $k(Z,Z) = U \operatorname{diag}(\lambda_1,\dots,\lambda_\ell)U^\top$ with an orthogonal matrix $U = [u_1,\dots,u_\ell]$ and $\lambda_1 \geq \cdots \geq \lambda_\ell \geq 0$. Note that

$$k_s^Z(x,y) = \sum_{i=1}^{s} \mathbf{1}_{\{\lambda_i>0\}} \frac{1}{\lambda_i}(u_i^\top k(Z,x))(u_i^\top k(Z,y)) \tag{5.6}$$

is actually a truncated Mercer decomposition of k^Z with regard to the measure $\mu_Z = \frac{1}{\ell}\sum_{i=1}^{\ell}\delta_{z_i}$, since

$$\begin{aligned}&\left\langle u_i^\top k(Z,\cdot), u_j^\top k(Z,\cdot)\right\rangle_{L^2(\mu_Z)}\\ &= \frac{1}{\ell}u_i^\top k(Z,Z)k(Z,Z)u_j = \frac{\lambda_i\lambda_j}{\ell}\delta_{ij}.\end{aligned}$$

This fact is at the heart of our analysis: k_s^Z is 'optimal' s-rank approximation for the measure μ_Z, and the statistical learning theory connects estimates in empirical measure and the original measure.

Let us denote by $P_{Z,s} : \mathcal{H}_k \to \mathcal{H}_k$ the linear operator given by $k(\cdot,x) \mapsto k_s^Z(\cdot,x)$ for all $x \in \mathcal{X}$. We shall also simply write $P_Z = P_{Z,\ell}$.

Lemma 5.1. *$P_{Z,s}$ is an orthogonal projection in $\mathcal{H}$.*

This projection is related to the quantity of interest, in that

$$k_s^Z(x,x) = \langle k(\cdot,x), P_{Z,s}k(\cdot,x)\rangle_{\mathcal{H}_k} = \|P_{Z,s}k(\cdot,x)\|_{\mathcal{H}_k}^2.$$

Thus, we have $k(x,x) - k_s^Z(x,x) = \|P_{Z,s}^\perp k(\cdot,x)\|_{\mathcal{H}_k}^2$ by using $P_{Z,s}^\perp$, the orthogonal complement of $P_{Z,s}$. So we are now interested in estimating the integral

$\mu(\sqrt{k - k_s^Z}) = \int_{\mathcal{X}} \|P_{Z,s}^{\perp} k(\cdot, x)\|_{\mathcal{H}_k} \,\mathrm{d}\mu(x)$ from the viewpoint of the projection operator. We first estimate its empirical counterpart $\mu_Z(\sqrt{k - k_s^Z}) = \frac{1}{\ell}\sum_{i=1}^{\ell} \|P_{Z,s}^{\perp} k(\cdot, z_i)\|_{\mathcal{H}_k}$, where $\mu_Z = \frac{1}{\ell}\sum_{i=1}^{\ell} \delta_{z_i}$ is the empirical measure. Indeed, we have the following identity regarding $\mu_Z(k - k_s^Z)$:

Lemma 5.2. *For any ℓ-point sample $Z \subset \mathcal{X}$, we have*

$$\mu_Z(\sqrt{k - k_s^Z})^2 \le \mu_Z(k - k_s^Z) = \frac{1}{\ell}\sum_{i=s+1}^{\ell} \lambda_i$$

where $\lambda_1 \ge \cdots \ge \lambda_\ell$ are eigenvalues of $k(Z, Z)$.

When Z is given by an i.i.d. sampling, the decay of eigenvalues λ_i enjoys the rapid decay given by σ_i in the following sense:

Lemma 5.3. *Let $Z = (z_i)_{i=1}^{\ell}$ be an ℓ-point independent sample from μ. Then, for the eigenvalues $\lambda_1 \ge \cdots \ge \lambda_\ell$ of $k(Z, Z)$, we have*

$$\mathbb{E}\left[\frac{1}{\ell}\sum_{i=s+1}^{\ell} \lambda_i\right] \le \sum_{i>s} \sigma_i.$$

For a general random orthogonal projection operator, we can prove the following bound by using arguments in statistical learning theory (Section 5.A). Recall from the previous chapter that we have defined $k_{\max} := \sup_{x\in\mathcal{X}} k(x, x)$.

Theorem 5.4. *Let $Z = (z_i)_{i=1}^{\ell}$ be an ℓ-point independent sample from μ and P be a random orthogonal projection in $\mathcal{H}_k$ possibly depending on Z. For any integer $m \ge 1$, we have the following bound:*

$$\begin{aligned}\mathbb{E}\left[\int_{\mathcal{X}} \|Pk(\cdot, x)\|_{\mathcal{H}_k} \,\mathrm{d}\mu(x)\right] &\le \mathbb{E}\left[\frac{2}{\ell}\sum_{i=1}^{\ell} \|Pk(\cdot, z_i)\|_{\mathcal{H}_k}\right] \\ &\quad + 4\sqrt{\sum_{i>m} \sigma_i + \frac{\sqrt{k_{\max}}}{\ell}\left(\frac{80m^2 \log(1 + 2\ell)}{9} + 69\right)},\end{aligned}$$

where the expectation is taken regarding the draws of Z.

Recall that $\mu(\sqrt{k - k_s^Z}) = \int_{\mathcal{X}} \|P_{Z,s}^{\perp} k(\cdot, x)\|_{\mathcal{H}_k} \,\mathrm{d}\mu(x)$. By combining this theorem when $P = P_{Z,s}^{\perp}$ and Lemma 5.2 & 5.3, we can obtain the following:

Corollary 5.5. *Let $Z = (z_i)_{i=1}^{\ell}$ be an ℓ-point independent sample from μ. Then, for any integer $m \geq 1$, we have*

$$\mathbb{E}\Big[\mu(\sqrt{k - k_s^Z})\Big] \leq 2\sqrt{\sum_{i>s} \sigma_i} + 4\sqrt{\sum_{i>m} \sigma_i} \\ + \frac{\sqrt{k_{\max}}}{\ell}\left(\frac{80m^2 \log(1 + 2\ell)}{9} + 69\right).$$

Remark 5.1. When $\sigma_j \lesssim e^{-\beta i^{1/d}}$ with a constant $\beta > 0$ and a positive integer d [typical for d-dimensional Gaussian kernel, see, e.g., 1, Section A.2], by taking $m \sim (\log \ell)^d$, we have a bound

$$\mathbb{E}\Big[\mu(\sqrt{k - k_s^Z})\Big] = \mathcal{O}\left(\sqrt{\sum_{i>s} \sigma_i} + \frac{(\log \ell)^{2d+1}}{\ell}\right)$$

for $\ell \geq 3$; see Appendix 5.B.6 for the proof. Since $k - k_s^Z \leq \sqrt{k_{\max}}\sqrt{k - k_s^Z}$, the same estimate applies to $\mathbb{E}[\mu(\sqrt{k - k_s^Z})]$. These also lead to an $(s + 1)$-point randomized convex kernel quadrature Q_{s+1} with the same order of $\mathbb{E}[\mathrm{wce}(Q_{s+1})]$. See Section 5.5 for details.

5.4 A refined low-rank approximation with general Z

The process of obtaining a good approximation k_{app} of k using k^Z can be decomposed into two parts:

$$k - k_{\mathrm{app}} = \underbrace{k - k^Z}_{\mathrm{A}} + \underbrace{k^Z - k_{\mathrm{app}}}_{\mathrm{B}}.$$

In the previous section, we have analyzed the case Z is i.i.d. and $k_{\mathrm{app}} = k_s^Z$. However, we can consider more general Z, and indeed we actually have a better way to select a subspace (i.e., k_{app}) from the finite-rank kernel k^Z rather than just using k_s^Z.

5.4.1 Part A: Estimating the error of k^Z for general Z

This part is relatively well-studied. Indeed, $\mu(k-k^Z) = \int_{\mathcal{X}}(k(x,x)-k^Z(x,x))\,\mathrm{d}\mu(x)$ for some non-i.i.d. Z can be bounded by using the results of weighted kernel quadrature. For example, Belhadji et al. [16] consider the worst-case error for the weighted integral

$$\mu(fg) = \int_{\mathcal{X}} f(x)g(x)\,\mathrm{d}\mu(x) \approx \sum_{i=1}^{\ell} w_i f(z_i) \tag{5.7}$$

for any $\|f\|_{\mathcal{H}_k} \leq 1$ and a fixed $g \in L^2(\mu)$ with $Z = (z_i)_{i=1}^{\ell}$ following a certain DPP. Now consider the optimal worst-case error in the above approximation for the fixed point configuration Z:

$$\begin{aligned}\inf_{w_i} \sup_{\|f\|_{\mathcal{H}_k}\leq 1} \left|\mu(fg) - \sum_{i=1}^{\ell} w_i f(z_i)\right| &= \sup_{\|f\|\leq 1} \left|\left\langle f, \int_{\mathcal{X}} k(\cdot,x)g(x)\,\mathrm{d}\mu(x) - \sum_{i=1}^{\ell} w_i k(\cdot, z_i)\right\rangle_{\mathcal{H}_k}\right| \\ &= \inf_{w_i} \left\|\mathcal{K}g - \sum_{i=1}^{\ell} w_i k(\cdot, z_i)\right\|_{\mathcal{H}_k} = \|P_Z^{\perp}\mathcal{K}g\|_{\mathcal{H}_k}.\end{aligned} \tag{5.8}$$

By using this, we can prove the following estimate:

Proposition 5.6. *For any finite subset $Z \subset \mathcal{X}$ and any integer $m \geq 0$, we have*

$$\mu(k - k^Z) = \sum_{i=1}^{\infty} \|P_Z^{\perp}\mathcal{K}e_i\|_{\mathcal{H}_k}^2 \leq \sum_{i=1}^{m} \|P_Z^{\perp}\mathcal{K}e_i\|_{\mathcal{H}_k}^2 + \sum_{i>m} \sigma_i$$

where $(\sigma_i, e_i)_{i=1}^{\infty}$ are the eigenpairs of $\mathcal{K}$.

The papers Belhadji et al. [16, 17], Belhadji [15] give bounds on the worst-case error of the weighted kernel quadrature (5.8) when Z is given by some correlated sampling, whereas Bach [6] gives another bound when Z is given by an optimized weighted sampling rather than sampling from μ. By using (5.8) and Proposition 5.6, we can import their bounds on weighted kernel quadrature with non-i.i.d. Z to the estimate of $\mu(k-k^Z) = \int_X \|P_Z^{\perp}k(\cdot,x)\|_{\mathcal{H}_k}^2\,\mathrm{d}\mu(x)$. Here, we just give one such example:

Corollary 5.7. *Let $Z = (z_i)_{i=1}^\ell$ be taken from a DPP given by the projection kernel $p(x, y) = \sum_{i=1}^\ell e_i(x)e_i(y)$ with a reference measure μ, i.e., $\mathbb{P}(Z \in A) = \frac{1}{\ell!}\int_A \det p(Z, Z)\,\mathrm{d}\mu^{\otimes\ell}(Z)$ for any Borel set $A \subset \mathcal{X}^d$. Then, for any integer $m \geq 0$, we have*

$$\mathbb{E}\big[\mu(k - k^Z)\big] \leq \sum_{i>m} \sigma_i + 4m \sum_{i>\ell} \sigma_i,$$

where the expectation is taken regarding the draws of Z.

In any case, by using those non-i.i.d. points, we can obtain a better Z in the sense that $\int_{\mathcal{X}}(k(x, x) - k^Z(x, x))\,\mathrm{d}\mu(x)$ attains a sharper upper bound than the bound given in the previous section for an ℓ-point i.i.d. sample from μ. However, for a general Z, it is not necessary sensible to execute the SVD of $k(Z, Z)$ and get k_s^Z accordingly, as an SVD of $k(Z, Z)$ corresponds to approximating μ by the empirical measure $\frac{1}{\ell}\sum_{i=1}^\ell \delta_{z_i}$ (indeed, this observation is the key to the results in the previous section). Thus, for points Z not given by i.i.d. sampling, there should exist a better choice of k_{app} than k_s^Z. We discuss this in the following section.

5.4.2 Part B: Mercer decomposition of k^Z

Instead of using k_s^Z, we propose to compute the Mercer decomposition of k^Z with respect to μ and truncate it to get $k_{s,\mu}^Z$, which is defined in the following. This is doable if we have knowledge of $h_\mu(x, y) := \int_{\mathcal{X}} k(x, t)k(t, y)\,\mathrm{d}\mu(t)$, since k^Z is a finite-dimensional kernel. We can prove the following:

Lemma 5.8. *We have $h_\mu(x, y) = \sum_{i=1}^\infty \sigma_i^2 e_i(x)e_i(y)$.*

We now discuss how h_μ can be used to derive the Mercer decomposition of k^Z. Note that this can be regarded as a generalization of Santin and Schaback [146, Section 6]. Let $\mathcal{K}^Z : L^2(\mu) \to L^2(\mu)$ be the integral operator given by $g \mapsto \int_{\mathcal{X}} k^Z(\cdot, x)g(x)\,\mathrm{d}\mu(x)$.

For functions of the form $f = a^\top k(Z, \cdot)$ and $g = b^\top k(Z, \cdot)$ with $a, b \in \mathbb{R}^\ell$, we have

$$\langle f, g\rangle_{L^2(\mu)} = \int_{\mathcal{X}} a^\top k(Z, x)k(x, Z)b\,\mathrm{d}\mu(x) = a^\top h_\mu(Z, Z)b. \tag{5.9}$$

So, if we write $h_\mu(Z, Z) = H^\top H$ by using an $H \in \mathbb{R}^{\ell\times\ell}$ (since $h_\mu(Z, Z)$ is positive semi-definite), an element $f = a^\top k(Z, \cdot) \in L^2(\mu)$ is non-zero if and only if $Ha \neq 0$. Furthermore, we have

$$\begin{aligned}\mathcal{K}^Z f &= \int_{\mathcal{X}} k(\cdot, Z)k(Z, Z)^+ k(Z, x)k(x, Z)a \,\mathrm{d}\mu(x)\\ &= k(\cdot, Z)k(Z, Z)^+ h_\mu(Z, Z)a\\ &= \left[k(Z, Z)^+ h_\mu(Z, Z)a\right]^\top k(Z, \cdot). \qquad (5.10)\end{aligned}$$

Thus, f is a nontrivial eigenfunction of $\mathcal{K}^Z$, if $Ha \neq 0$ and a is an eigenvector of $k(Z, Z)^+ h_\mu(Z, Z)$. It is equivalent to $c = Ha$ being an eigenvector of $Hk(Z, Z)^+ H^\top$.

Let us decompose this matrix by SVD as $Hk(Z, Z)^+ H^\top = V \operatorname{diag}(\kappa_1, \dots, \kappa_\ell)V^\top$, where the $V = [v_1, \dots, v_\ell] \in \mathbb{R}^{\ell\times\ell}$ is an orthogonal matrix and $\kappa_1 \geq \cdots \geq \kappa_\ell \geq 0$. Then, we have

$$Hk(Z, Z)^+ H^\top = \sum_{i=1}^{\ell} \kappa_i v_i v_i^\top.$$

Let us consider $f_i = (H^+ v_i)^\top k(Z, \cdot) = v_i^\top (H^+)^\top k(Z, \cdot)$ for $i = 1, \dots, \ell$ as candidates of eigenfunctions of $\mathcal{K}^Z$. We can actually prove the following:

Lemma 5.9. *The set $\{f_i \mid i \geq 1,\, \kappa_i > 0\}$ forms an orthonormal subset of $L^2(\mu)$ whose elements are eigenfunctions of $\mathcal{K}^Z$.*

Let us define $k_\mu^Z(x, y) := \sum_{i=1}^{\ell} \kappa_i f_i(x) f_i(y)$; note that this is computable. From the above lemma, this expression is a natural candidate for "Mercer decomposition" of k^Z. We can prove that it actually coincides with $k^Z(x, y)$ μ-almost everywhere, and so the decomposition is independent of the choice of H up to μ-null sets:

Proposition 5.10. *There exists a measurable set $A \subset \mathcal{X}$ depending on Z with $\mu(A) = 1$ such that $k^Z(x, y) = k_\mu^Z(x, y)$ holds for all $x, y \in A$. Moreover, we can take $A = \mathcal{X}$ if $\ker h_\mu(Z, Z) \subset \ker k(Z, Z)$.*

Now we just define $k_{s,\mu}^Z$ for $s \leq \ell$ as follows:

$$k_{s,\mu}^Z(x, y) := \sum_{i=1}^{s} \kappa_i f_i(x) f_i(y). \qquad (5.11)$$

Theorem 5.11. *We have $\mu(k_\mu^Z - k_{s,\mu}^Z) \le \sum_{i=s+1}^{\ell} \sigma_i$ for any $Z = (z_i)_{i=1}^{\ell} \subset \mathcal{X}$.*

Proof. The left-hand side is equal to $\sum_{i=s+1}^{\ell} \kappa_i$ from Lemma 5.9 and the definition of the kernels. Thus, it suffices to prove $\kappa_i \le \sigma_i$ for each i. It directly follows from the min-max principle (or Weyl's inequality) as $k - k_\mu^Z$ is positive definite on an $A \subset \mathcal{X}$ with $\mu(A) = 1$ from Proposition 5.10. □

Remark 5.2. The choice of the matrix H with $H^\top H = h_\mu(Z, Z)$ does not affect the theory but might affect the numerical errors. We have used the matrix square-root $h_\mu(Z, Z)^{1/2}$, i.e., the symmetric and positive semi-definite matrix H with $H^2 = h_\mu(Z, Z)$, throughout the experiments in Section 5.5, so that we just need to take the pseudo-inverse of positive semi-definite matrices.

Approximate Mercer decomposition. When we have no access to the function h_μ, we can just approximate it by using an empirical measure. For a $X = (x_j)_{j=1}^M \subset \mathcal{X}$, denote by h_X the function given by replacing μ in h_μ with the empirical measure with points X:

$$h_X(x, y) = \frac{1}{M} \sum_{j=1}^{M} k(x, x_j) k(x_j, y) = \frac{1}{M} k(x, X) k(X, y).$$

We can actually replace every h_μ by h_X in the above construction to define k_X^Z and $k_{s,X}^Z$. This approximation is already mentioned by Santin and Schaback [146] without theoretical guarantee. Another remark is that when restricted on the set X, it is equivalent to the best s-rank approximation of $k^Z(X, X)$ in the Gram-matrix case [166, 177], since the L^2-norm for the uniform measure on X just corresponds to the ℓ^2-norm in $\mathbb{R}^{|X|}$.

Note that we have $k_X^Z(X, X) = k^Z(X, X)$ from Proposition 5.10 in the discrete case. As we have $\ker h_X(Z, Z) = \ker k(Z, X) k(X, Z) = \ker k(X, Z)$, we additionally obtain the following sufficient condition from Proposition 5.10.

Proposition 5.12. *$k_X^Z(x, y) = k^Z(x, y)$ holds for all $x, y \in X$. Moreover, if $\ker k(X, Z) \subset \ker k(Z, Z)$, then we have $k_X^Z = k^Z$ over the whole $\mathcal{X}$.*

In particular, we have $k_X^Z = k^Z$ whenever $Z \subset X$. These (at least μ-a.s.) equalities given in Proposition 5.10 & 5.12 are necessary for the applications to kernel quadrature, since we need $k - k_{\text{app}}$ to be positive definite for exploiting the existing guarantees such as Theorem 5.15 in the next section.

Although checking $k_X^Z = k^Z$ is not an easy task, from the first part of Proposition 5.12, $k_{s,X}^Z$ satisfies the following estimate in terms of the empirical measure μ_X.

Proposition 5.13. *Let $Z \subset \mathcal{X}$ be a fixed subset and X be an M-point independent sample from μ. Then, we have*

$$\mathbb{E}\bigl[\mu_X(k^Z - k_{s,X}^Z)\bigr] = \mathbb{E}\bigl[\mu_X(k_X^Z - k_{s,X}^Z)\bigr] \le \sum_{i>s} \sigma_i,$$

where the expectation is taken regarding the draws of X.

We can also give a bound of the resulting error $\mu(k^Z - k_{s,X}^Z)$ again by using the arguments from learning theory, but under an additional assumption as stated in the following. Nevertheless, Proposition 5.13 is already sufficient for our application in kernel quadrature; see Theorem 5.16.

Proposition 5.14. *Under the same setting as in Proposition 5.13, if* $\ker k(X, Z) \subset \ker k(Z, Z)$ *holds almost surely for the draws of X, we have*

$$\mathbb{E}\Bigl[\mu(\sqrt{k^Z - k_{s,X}^Z})\Bigr] \le 2\sqrt{\sum_{i>M} \sigma_i} + 4\sqrt{\sum_{i>m} \sigma_i} + \frac{\sqrt{k_{\max}}}{M}\left(\frac{80m^2 \log(1 + 2M)}{9} + 69\right).$$

for any integer $m \ge 1$.

Remark 5.3. The assumption $\ker k(X, Z) \subset \ker k(Z, Z)$ seems to be very hard to check in practice. An example with this property is $(\mathcal{X}, k, \mu)$ such that $\mathcal{X} = \mathbb{R}^D$ with $D, M > \ell$, the kernel k is just the Euclidean inner product on $\mathbb{R}^D$, and μ is given by a Gaussian distribution with a nonsingular covariance matrix.

This being said, we have some ways to avoid this issue in practice. One way is to use $X \cup Z$ instead of X so that the condition automatically holds. Then, the above order of estimate should still hold when $\ell \ll M$, though it complicates the analysis. Another way is effective when we use k_X^Z for constructing a kernel quadrature from an empirical measure given by X itself; see the next section for details.

5.5 Application to kernel quadrature

Let us give error bounds for kernel quadrature as a consequence of the previous sections. We are mainly concerned with the kernel quadrature of the form (5.7) without the weight function, i.e., the case when $g = 1$, for efficiently discretizing the probability measure μ.

Given an n-point quadrature rule $Q_n : f \mapsto \sum_{i=1}^n w_i f(x_i)$ with weights $w_i \in \mathbb{R}$ and points $x_i \in \mathcal{X}$, the worst-case error of Q_n with respect to the RKHS $\mathcal{H}_k$ and the target measure μ is defined as $\mathrm{wce}(Q_n; \mathcal{H}_k, \mu) = \mathrm{MMD}_k(Q_n, \mu) = \sup_{\|f\|_{\mathcal{H}_k} \le 1} |Q_n(f) - \mu(f)|$ as explained in Section 1.2. We again call Q_n *convex* if it forms a probability measure, i.e., $w_i \ge 0$ and $\sum_{i=1}^n w_i = 1$.

Suppose we are given an s-rank kernel approximation $k_{\mathrm{app}}(x, y) = \sum_{i=1}^s c_i \varphi_i(x) \varphi_i(y)$ with $c_i \ge 0$ and $k - k_{\mathrm{app}}$ being positive definite (μ-almost surely). The following is taken from the previous chapter (Theorem 4.6 & 4.9 combined).

Theorem 5.15. *If an n-point convex quadrature Q_n satisfies $Q_n(\varphi_i) = \mu(\varphi_i)$ for $1 \le i \le s$ and $Q_n(\sqrt{k - k_{\mathrm{app}}}) \le \mu(\sqrt{k - k_{\mathrm{app}}})$, then we have*

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu) \le 2\mu(\sqrt{k - k_{\mathrm{app}}}).$$

Moreover, such a quadrature Q_n exists with $n = s + 1$.

Although there is a randomized algorithm for constructing the Q_n stated in the above theorem (Algorithm 4.2 with modification), it has two issues; it requires exact values of $\mu(\varphi_i)$ (and $\mu(\sqrt{k - k_{\mathrm{app}}})$) and its computational complexity has no useful upper bound unless we have additional assumptions such as well-behaved moments of test functions φ_i or structure like a product kernel with a product measure as in Chapter 3. This being said, we can deduce updated convergence results for outputs of the algorithm as in Remark 5.1.

5.5.1 Kernel recombination

Instead of considering an "exact" quadrature, what we do in practice in this low-rank approach is matching the integrals against a large empirical measure [see also

1, Section 6], say $\mu_Y = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$ with $Y = (y_i)_{i=1}^N$. If we have

$$\begin{cases} Q_n(\varphi_i) = \mu_Y(\varphi_i), \quad 1 \le i \le s, \\ Q_n(\sqrt{k - k_{\mathrm{app}}}) \le \mu_Y(\sqrt{k - k_{\mathrm{app}}}), \end{cases} \tag{5.12}$$

then, from Theorem 5.15 with a target measure μ_Y and the triangle inequality of MMD, we have

$$\begin{aligned} \mathrm{wce}(Q_n; \mathcal{H}_k, \mu) &\le \mathrm{MMD}_k(Q_n, \mu_Y) + \mathrm{MMD}_k(\mu_Y, \mu) \\ &\le 2\mu_Y(\sqrt{k - k_{\mathrm{app}}}) + \mathrm{MMD}_k(\mu_Y, \mu). \end{aligned} \tag{5.13}$$

Indeed, such a quadrature Q_n with $n = s + 1$ points given by a subset of Y can be constructed via an algorithm called *recombination* [103, 163, 34].

Existing approaches with this kernel recombination have then been using an approximation k_{app} typically given by k_s^Z whose randomness is independent of the sample Y, but it is not a necessary requirement as long as we can expect an efficient bound of $\mu_Y(\sqrt{k - k_{\mathrm{app}}})$ in some sense. Another small but novel observation is that $k - k_{\mathrm{app}}$ being positive definite is only required on the sample Y in deriving the estimate (5.13); not over the support of μ in contrast to Theorem 5.15. These observations circumvent the issues mentioned in Remark 5.3 when using $k_{\mathrm{app}} = k_Y^Z$ ($k_{s,X}^Z$ with $X = Y$).

Let us now denote the kernel recombination in the form of a function as $Q_n = \mathrm{KQuad}(k_{\mathrm{app}}, Y)$, where the output Q_n is an n-point convex quadrature satisfying $n = s + 1$ and (5.12); note that the constraint is slightly different from what is given in Algorithm 4.1, but we can achieve (5.12) by replacing $k_{1,\mathrm{diag}}$ with $\sqrt{k_{1,\mathrm{diag}}}$ in the cited algorithm.

We can now prove the performance of low-rank approximations given in the previous section. Indeed, $k_{s,Y}^Z$ and $k_{s,\mu}^Z$ with $s = n - 1$ have the following same estimate.

Theorem 5.16. *Let $Z \subset \mathcal{X}$ be a fixed subset and Y be an N-point independent sample from μ. The random convex quadrature $Q_n = \mathrm{KQuad}(k_{n-1,Y}^Z, Y)$ satisfies*

$$\mathbb{E}[\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)] \le 2\mu(\sqrt{k - k^Z}) + 2\sqrt{\sum_{i \ge n} \sigma_i} + \sqrt{\frac{c_{k,\mu}}{N}}, \tag{5.14}$$

where $c_{k,\mu} := \mu(k) - \iint_{\mathcal{X}\times\mathcal{X}} k(x, y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)$ and the expectation is taken regarding the draws of Y. The estimate (5.14) holds also for $Q_n = \mathrm{KQuad}(k_{n-1,\mu}^Z, Y)$.

5.5.2 Numerical examples

In this section, we compare the numerical performance of $k_{s,Y}^Z$ and $k_{s,\mu}^Z$ for kernel quadrature with the conventional Nyström approximation for a non-i.i.d. Z in the setting that we can explicitly compute the worst-case error.

Periodic Sobolev spaces. The class of RKHS we use is called periodic Sobolev spaces of functions on $\mathcal{X} = [0, 1]$ (a.k.a. Korobov spaces), and given by the following kernel for a positive integer r:

$$k_r(x, y) = 1 + \frac{(-1)^{r-1}(2\pi)^{2r}}{(2r)!} B_{2r}(|x - y|),$$

where B_{2r} is the $2r$-th Bernoulli polynomial [173, 6]. We consider the case μ being the uniform measure, where the eigenfunctions of the integral operator $\mathcal{K}$ are known to be $1, \sqrt{2}\cos(2\pi m\,\cdot), \sqrt{2}\sin(2\pi m\,\cdot)$ with eigenvalues respectively $1, m^{-2r}, m^{-2r}$ for each positive integer m. This RKHS is commonly used for measuring the performance of kernel quadrature methods [82, 6, 16]. We also consider its products: $k_r^{\otimes d}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^d k_r(x_i, y_i)$ and μ being the uniform measure on the hypercube $\mathcal{X} = [0, 1]^d$.

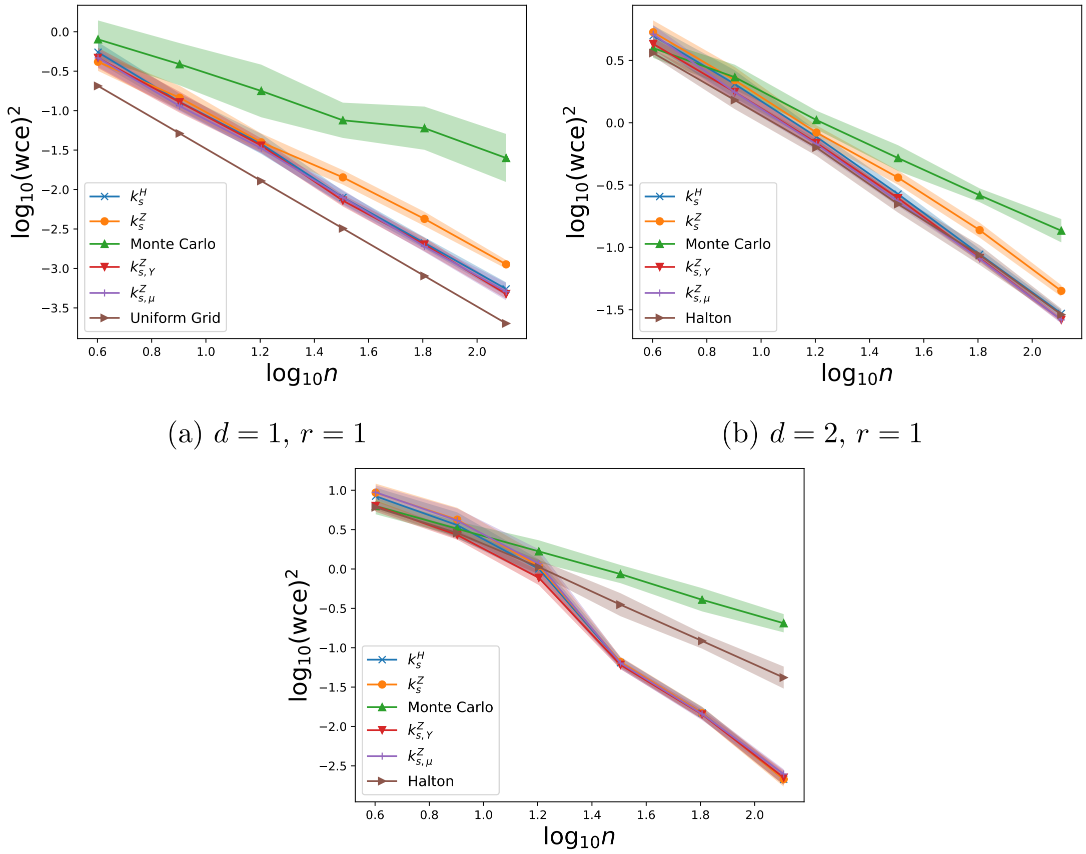
By considering the eigenvalues, we can see that $h_\mu = k_{2r}^{\otimes d}$ for each kernel $k_r^{\otimes d}$ from Remark 5.8.

Experiments. In the experiments for the kernel $k_r^{\otimes d}$, we compared the worst-case error of n-point kernel quadrature rules given by $Q_n = \mathrm{KQuad}(k_{\mathrm{app}}, Y)$ with $k_{\mathrm{app}} = k_s^H, k_s^Z, k_{s,Y}^Z, k_{s,\mu}^Z$ $(s = n - 1)$ under the following setting:

- Y is an N-point independent sample from μ with $N = n^2$ (Figure 5.1) or $N = n^3$ (Figure 5.2).
- H is the uniform grid $\{i/n \mid i = 1, \dots, n\}$ $(d = 1)$ or the Halton sequence with Owen scrambling [64, 132] $(d \geq 2)$.
- Z is the union of H and another $20n$-point independent sample from $\nu^{\otimes d}$, where ν is the 1-dimensional $(2, 5)$-Beta distribution, whose density is proportional to $x(1 - x)^4$ for $x \in [0, 1]$.

We additionally compared ‘**Monte Carlo**’: uniform weights $1/n$ with i.i.d. sample $(x_i)_{i=1}^n$ from μ, ‘**Uniform Grid**’ ($d = 1$): points in H with uniform weights $1/n$ (known to be *optimal* for each n), and ‘**Halton**’ ($d \geq 2$): points in an independent copy of H with uniform weights $1/n$.

The aim of this experiment was to see if the proposed methods ($k_{s,Y}^Z$ and $k_{s,\mu}^Z$) can actually recover a ‘good’ subspace of the RKHS given by k^Z with Z not summarizing μ. To do so, we mixed H (a ‘good’ summary of μ) and an i.i.d. sample from ν to determine Z.



(a) $d = 1,\ r = 1$

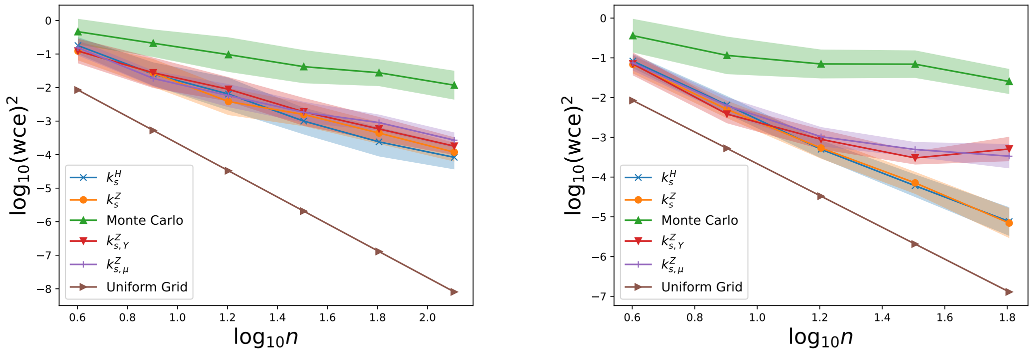
(b) $d = 2,\ r = 1$

(c) $d = 3,\ r = 3$

Figure 5.1: Experiments in periodic Sobolev spaces with reproducing kernel $k_r^{\otimes d}$. Average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 20 samples plotted with their standard deviation.

Figure 5.1 shows the results for $(d, r) = (1, 1), (2, 1), (3, 3)$ with $N = n^2$ and

$n = 4, 8, 16, 32, 64, 128$. From Figure 5.1(a, b), we can see that our methods indeed recover (and perform slightly better than) the rate of k^H from a contaminated sample Z. In Figure 5.1(c), the four low-rank methods all perform equally well, and it seems that the dominating error is given by the term caused by $\mathrm{MMD}_k(\mu_Y, \mu)$.



(a) $N = n^2$

(b) $N = n^3$

Figure 5.2: Experiments in k_2 with $N = n^2, n^3$ for recombination algorithms. Average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 20 samples plotted with their standard deviation.

Figure 5.2 shows the results for $(d, r) = (1, 2)$ with $N = n^2$ or $N = n^3$ and $n = 4, 8, 16, 32, 64$. In this case, we can see that $k_{s,Y}^Z$ or $k_{s,\mu}^Z$ eventually suffers from numerical instability, which is also reported by Santin and Schaback [146]. Since their error inflation is not completely hidden even in the case $N = n^2$ unlike the previous experiments, one possible reason for the instability is that taking the pseudo-inverse of $k(Z, Z)$ or $h_\mu(Z, Z)^{1/2}$ in the algorithm becomes highly unstable when the spectral decay is fast. Although they have preferable guarantees in theory, its numerical error seems to harm the overall efficiency, and this issue needs to be addressed e.g. by circumventing the use of pseudo-inverse in future work.

Remark 5.4. Unlike the kernel quadrature with $k_{s,\mu}^Z$ or $k_{s,Y}^Z$, that with k_s^Z does not suffer from a similar numerical instability despite the use of $k(Z, Z)_s^+$. This phenomenon can be explained by the nature of Algorithm 4.1; it only requires (stable) *test functions* $\varphi_i = u_i^\top k(Z, \cdot)$ $(i = 1, \ldots, s)$ for its equality constraints, where u_i is the i-th eigenvector of $k(Z, Z)$, while the (possibly unstable) diagonal term

$k_s^Z(x,x)$ appears in the inequality constraint, which can empirically be omitted (see Section 4.E.2).

Computational complexity. By letting ℓ, N (larger than s) respectively be the cardinality of Z and Y, we can express the computational steps of KQuad(k_{app}, Y) with $k_{\mathrm{app}} = k_s^Z, k_{s,Y}^Z, k_{s,\mu}^Z$ as follows:

- Using k_s^Z takes $\mathcal{O}(s\ell N + s\ell^2 + s^3 \log(N/s))$, but by omitting the (empirically unnecessary) inequality constraint, it can be reduced to $\mathcal{O}(\ell N + s\ell^2 + s^3 \log(N/s))$ (see Remark 4.2).
- Using $k_{s,Y}^Z$ takes $\mathcal{O}(\ell^3 + \ell^2 N + s^3 \log(N/s))$, where $\mathcal{O}(\ell^3)$ and $\mathcal{O}(\ell^2 N)$ respectively come from computing $k(Z,Z)^+$ and $h_Y(Z,Z)$.
- Using $k_{s,\mu}^Z$ takes $\mathcal{O}(\ell^3 + s\ell N + s^3 \log(N/s))$ (if h_μ available), where $\mathcal{O}(\ell^3)$ is from computing $k(Z,Z)^+$.

For example, in the case of Figure 5.1(c) with $n = 128$, the average time per trial was respectively 26.5, 226, 216 seconds for $k_s^Z, k_{s,Y}^Z, k_{s,\mu}^Z$, while it was 52.6, 57.8, 41.2 seconds for the case of Figure 5.2(b) with $n = 64$.[1]

5.6 Concluding remarks

In this chapter, we have studied the performance of several Nyström-type approximations k_{app} of a positive definite kernel k associated with a probability measure μ, in terms of the error $\mu(\sqrt{k - k_{\mathrm{app}}})$. We first improved the bounds for k_s^Z, the conventional Nyström approximation based on an i.i.d. Z and the use of SVD, by leveraging results in statistical learning theory. We then went beyond the i.i.d. setting and considered general Z including DPPs; we further introduced two competitors of k_s^Z, i.e., $k_{s,\mu}^Z$ and $k_{s,X}^Z$, which are given by directly computing the Mercer decomposition of the finite-rank kernel k^Z against the measure μ and the empirical measure μ_X, respectively. Finally, we used our results to improve the

[1] All the experiments were conducted on a MacBook Pro with Apple M1 Max chip and 32GB unified memory. Code is available at the `nystrom` folder in `https://github.com/satoshi-hayakawa/kernel-quadrature`.

theoretical guarantees for convex kernel quadrature introduced in Chapter 4, and provided numerical results to illustrate the difference between the conventional k_s^Z and the newly proposed $k_{s,\mu}^Z$ and $k_{s,X}^Z$.

Despite its nice theoretical properties, a limitation of our second contribution (i.e., the proposed kernel approximations) is that they involve the computation of a pseudo-inverse, which can be numerically unstable when there is a rapid spectral decay. This point should be addressed in future work, but one promising approach in the context of kernel quadrature is to conceptually learn from the stability of k_s^Z mentioned in Remark 5.4; if we see the construction of the low-rank kernel as optimization of the vectors u_i for which functions $u_i^\top k(Z, \cdot)$ well approximate $\mathcal{H}_{k^Z}$ in terms of $L^2(\mu)$ metric, we can possibly leverage the stability of convex optimization for instance.

Appendix for Chapter 5

5.A Tools from statistical learning theory

In this section, $\mathcal{F}$ always denotes a class of functions from $\mathcal{X}$ to $\mathbb{R}$, i.e., $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$. Let us define the Rademacher complexity of $\mathcal{F}$ with respect to the sample $Z = (z_i)_{i=1}^\ell \subset \mathcal{X}$ as follows [121, Definition 3.1]:

$$\mathcal{R}_Z(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{j=1}^{\ell} s_j f(z_j) \,\middle|\, Z \right]$$

where the conditional expectation is taken with regard to the Rademacher variables, i.e., i.i.d. variables s_j uniform in $\{\pm 1\}$.

The following is a version of the uniform law of large numbers, though we only use one side of the inequality.

Proposition 5.17 (121, Theorem 3.3)**.** *Let Z be an ℓ-point independent sample from μ. If there is a $B > 0$ such that $\|f\|_\infty \le B$ for every $f \in \mathcal{F}$, then with probability at least $1 - \delta$, we have*

$$\sup_{f \in \mathcal{F}} (\mu(f) - \mu_Z(f)) \le 2\mathbb{E}[\mathcal{R}_Z(\mathcal{F})] + \sqrt{\frac{2B^2}{\ell} \log \frac{1}{\delta}}.$$

For a pseudo metric d on $\mathcal{F}$, we denote the *ε-convering number* of $\mathcal{F}$ by $\mathcal{N}(\mathcal{F}, d; \varepsilon)$. Namely, $\mathcal{N}(\mathcal{F}, d; \varepsilon)$ is the infimum of positive integers N such that there exist $f_1, \dots, f_N \in \mathcal{F}$ satisfying $\min_{1\le i\le N} d(f_i, g) \le \varepsilon$ for all $g \in \mathcal{F}$.

Let us define a pseudo-metric $d_Z(f, g) := \sqrt{\frac{1}{\ell}\sum_{j=1}^{\ell}(f(z_i) - g(z_i))^2}$. The following assertion is a version of Dudley's integral entropy bound [158, Lemma A.3; see Srebro and Sridharan [157] for a correction of the constant].

Proposition 5.18 (Dudley integral). *For any ℓ-point sample $Z = (z_i)_{i=1}^{\ell} \subset \mathcal{X}$, we have*

$$\mathcal{R}_Z(\mathcal{F}) \le \frac{12}{\sqrt{\ell}} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, d_Z; \varepsilon)}\, \mathrm{d}\varepsilon.$$

The following is a straightforward modification of Schmidt-Hieber [148, Lemma 4] tailored to our setting. It originates from an analysis of empirical risk minimizers, and this kind of technique has also been known in earlier work under the name of local Rademacher complexities [62, 92, 53].

Proposition 5.19. *Let $\mathcal{F} \subset L^\infty(\mu)$ be a set of functions with $f \ge 0$ and $\|f\|_{L^\infty(\mu)} \le F$ for all $f \in \mathcal{F}$, where $F > 0$ is a constant. If $\hat{f}$ is a random function in $\mathcal{F}$ possibly depending on Z, then, for every $\varepsilon > 0$, we have*

$$\mathbb{E}\Big[\mu(\hat{f})\Big] \le 2\mathbb{E}\Big[\mu_Z(\hat{f})\Big] + \frac{F}{\ell}\left(\frac{80}{9}\log N + 64\right) + 5\varepsilon,$$

where $N := \max\{3, \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^1(\mu)}; \varepsilon)\}$.

Proof. The proof here essentially follows the original proof, where we re-compute the constants as the condition is slightly different; see also Hayakawa and Suzuki [67, Theorem 2.6] and its remark.

Let $Z' = (z'_1, \dots z'_\ell)$ be an independent copy of Z. Let $\mathcal{F}_\varepsilon$ be an ε-covering of $\mathcal{F}$ in $L^1(\mu)$ with the cardinality N and f^* be a random element of $\mathcal{F}_\varepsilon$ such that $\mu(|\hat{f} - f^*|) \le \varepsilon$. Then, we have

$$\left|\mathbb{E}\Big[\mu_Z(\hat{f})\Big] - \mathbb{E}\Big[\mu(\hat{f})\Big]\right| = \left|\mathbb{E}\left[\frac{1}{\ell}\sum_{i=1}^{\ell}(\hat{f}(z_i) - \hat{f}(z'_i))\right]\right| \le \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{\ell}(f^*(z_i) - f^*(z'_i))\right|\right] + 2\varepsilon \tag{5.15}$$

Define $T := \max_{f \in \mathcal{F}_\varepsilon} \frac{\sum_{i=1}^{\ell}(f(z_i) - f(z_i'))}{r(f)}$, where we let $r(f) := \max\{c\sqrt{\frac{\log N}{\ell}}, \sqrt{\mu(f)}\}$ for each $f \in \mathcal{F}_\varepsilon$ with a constant $c > 0$ fixed afterwards. Thus, we obtain

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{\ell}(f^*(z_i) - f^*(z_i'))\right|\right] \le \mathbb{E}\left[\frac{r(f^*)T}{\ell}\right] \le \frac{1}{2}\mathbb{E}\big[r(f^*)^2\big] + \frac{1}{2\ell^2}\mathbb{E}\big[T^2\big]. \tag{5.16}$$

The first term can be evaluated as

$$\mathbb{E}\big[r(f^*)^2\big] \le c^2\frac{\log N}{\ell} + \mathbb{E}[\mu(f^*)] \le c^2\frac{\log N}{\ell} + \mathbb{E}[\mu(\hat{f})] + \varepsilon. \tag{5.17}$$

For the second term, we first have

$$\sum_{i=1}^{\ell}\mathbb{E}\left[\left(\frac{f(z_i) - f(z_i')}{r(f)}\right)^2\right] \le \sum_{i=1}^{\ell}\mathbb{E}\left[\frac{f(z_i)^2 + f(z_i')^2}{r(f)^2}\right] \le 2F\ell, \qquad f \in \mathcal{F}_\varepsilon.$$

Since we have $|f(z_i) - f(z_i')|/r(f) \le 2F/r(f) \le 2F\frac{\sqrt{\ell}}{c\sqrt{\log N}}$ uniformly for $f \in \mathcal{F}_\varepsilon$, Bernstein's inequality combined with the union-bound yields

$$\mathbb{P}\big(T^2 \ge t\big) = \mathbb{P}\Big(T \ge \sqrt{t}\Big) \le 2N\exp\left(-\frac{t}{4F(\ell + \frac{\sqrt{\ell t}}{3c\sqrt{\log N}})}\right) \le 2N\exp\left(-\frac{3c\sqrt{\log N}}{8F\sqrt{\ell}}\sqrt{t}\right)$$

for $t \ge 9c^2\ell\log N$. Therefore, we have

$$\begin{aligned}\mathbb{E}\big[T^2\big] = \int_0^\infty \mathbb{P}\big(T^2 \ge t\big)\ \mathrm{d}t &\le 9c^2\ell\log N + \int_{9c^2\ell\log N}^{\infty} 2N\exp\left(-\frac{3c\sqrt{\log N}}{8F\sqrt{\ell}}\sqrt{t}\right)\ \mathrm{d}t \\ &= 9c^2\ell\log N + 4N\left(8F\ell + \frac{64F^2\ell}{9c^2\log N}\right)\exp\left(-\frac{9c^2\log N}{8F}\right)\end{aligned}$$

Let us now set $c = \sqrt{8F/9}$ so that $9c^2 = 8F$. Then, we obtain $\mathbb{E}[T^2] \le 8F\ell\log N + 64F\ell$ since $N \ge 3$ by assumption. By combining it with (5.15)–(5.17), we finally obtain

$$\left|\mathbb{E}\Big[\mu_Z(\hat{f})\Big] - \mathbb{E}\Big[\mu(\hat{f})\Big]\right| \le \frac{1}{2}\mathbb{E}\Big[\mu(\hat{f})\Big] + \frac{(\frac{40}{9}F\log N + 32F)}{\ell} + \frac{5}{2}\varepsilon,$$

from which the desired inequality readily follows. □

5.B Proofs

5.B.1 Properties of the pseudo-inverse

For a matrix $A \in \mathbb{R}^{m \times n}$, its Moore–Penrose pseudo-inverse A^+ [136] is defined as the unique matrix $X \in \mathbb{R}^{n \times m}$ that satisfies

$$AXA = A, \quad XAX = X, \quad (AX)^\top = AX, \quad (XA)^\top = XA.$$

It also satisfies that A^+A is the orthogonal projection onto the orthogonal complement of $\ker A$ (the range of $A^\top$), while AA^+ is the orthogonal projection onto the range of A [136, 153]. We use these general properties of A^+ throughout Section 5.B. See e.g. Drineas et al. [43] for the concrete construction of such a matrix.

5.B.2 Proof of Lemma 5.1

Proof. Recall that we have the SVD $k(Z, Z) = U \operatorname{diag}(\lambda_1, \dots, \lambda_\ell)U^\top$ with an orthogonal matrix $U = [u_1, \dots, u_\ell]$. and $\lambda_1 \ge \cdots \ge \lambda_\ell \ge 0$. By using this notation, we have

$$k_s^Z(x, y) = \sum_{\substack{1 \le j \le s \\ \lambda_j > 0}} \frac{1}{\lambda_j} (u_j^\top k(Z, x))(u_j^\top k(Z, y)). \tag{5.18}$$

If we denote by $Q_j : \mathcal{H}_k \to \mathcal{H}_k$ the projection onto $\operatorname{span}\{u_j^\top k(Z, \cdot)\}$, we have

$$\begin{aligned}(u_j^\top k(Z, x))(u_j^\top k(Z, y)) &= \left\langle u_j^\top k(Z, \cdot), k(\cdot, x) \right\rangle_{\mathcal{H}_k} \left\langle u_j^\top k(Z, \cdot), k(\cdot, y) \right\rangle_{\mathcal{H}_k} \\ &= \|u_j^\top k(Z, \cdot)\|_{\mathcal{H}_k}^2 \left\langle Q_j k(\cdot, x), Q_j k(\cdot, y) \right\rangle_{\mathcal{H}_k} \\ &= \lambda_j \left\langle Q_j k(\cdot, x), Q_j k(\cdot, y) \right\rangle_{\mathcal{H}_k}, \end{aligned} \tag{5.19}$$

where the last inequality follows from $\left\langle u_i^\top k(Z, \cdot), u_j^\top k(Z, \cdot) \right\rangle_{\mathcal{H}_k} = u_i^\top k(Z, Z)u_j = \delta_{ij}\lambda_j$. Now let $\widetilde{P}_{Z,s}$ be the orthogonal projection onto $\operatorname{span}\{u_j^\top k(Z, \cdot)\}_{j=1}^s$ in $\mathcal{H}_k$. We prove $\widetilde{P}_{Z,s} = P_{Z,s}$. From the orthogonality of $\{u_j^\top k(Z, \cdot)\}_{j=1}^s$ we have $\widetilde{P}_{Z,s} = \sum_{j=1}^s Q_j$ and

$$\begin{aligned}\left\langle k(\cdot, x), k_s^Z(\cdot, y) \right\rangle_{\mathcal{H}_k} = k_s^Z(x, y) &= \sum_{j=1}^s \left\langle Q_j k(\cdot, x), Q_j k(\cdot, y) \right\rangle_{\mathcal{H}_k} \\ &= \left\langle \widetilde{P}_{Z,s} k(\cdot, x), \widetilde{P}_{Z,s} k(\cdot, y) \right\rangle_{\mathcal{H}_k} = \left\langle k(\cdot, x), \widetilde{P}_{Z,s} k(\cdot, y) \right\rangle_{\mathcal{H}_k}\end{aligned}$$

for all $x, y \in \mathcal{X}$. In particular, $k_s^Z(\cdot, y) = \widetilde{P}_{Z,s} k(\cdot, y)$, so we have $\widetilde{P}_{Z,s} = P_{Z,s}$. □

5.B.3 Proof of Lemma 5.2

Proof. The inequality follows from Cauchy–Schwarz. Let us prove the equality.

We use the notation Q_j from the proof of Lemma 5.1. We first obtain $P_Z k(\cdot, z_i) = k(\cdot, z_i)$ for $i = 1, \dots, \ell$, since P_Z is a projection onto $\operatorname{span}\{k(\cdot, z_i)\}_{i=1}^{\ell}$. Thus, we have $P_{Z,s}^{\perp} k(\cdot, z_i) = (P_Z - P_{Z,s}) k(\cdot, z_i) = (Q_{s+1} + \cdots + Q_\ell) k(\cdot, z_i)$, and so

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \| P_{Z,s}^{\perp} k(\cdot, z_i) \|_{\mathcal{H}_k}^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} \sum_{\substack{s+1 \le j \le \ell \\ \lambda_j > 0}} \frac{1}{\lambda_j} (u_j^\top k(Z, z_i))^2$$

by using (5.19). Since $k(Z, Z) = U \operatorname{diag}(\lambda_1, \dots, \lambda_\ell) U^\top = \sum_{i=1}^{\ell} \lambda_i u_i u_i^\top$, we can explicitly calculate

$$u_j^\top k(Z, z_i) = u_j^\top \sum_{i=1}^{\ell} \lambda_i u_i u_i^\top \mathbf{1}_j = \lambda_j u_i^\top \mathbf{1}_j,$$

where $\mathbf{1}_j \in \mathbb{R}^\ell$ is the vector with 1 in the j-th coordinate and 0 in the other coordinates. As U is an $\ell \times \ell$ orthogonal matrix, we actually have $\sum_{i=1}^{\ell} (u_i^\top \mathbf{1}_j)^2 = 1$ for each $j = 1, \dots, \ell$.

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \sum_{\substack{s+1 \le j \le \ell \\ \lambda_j > 0}} \frac{1}{\lambda_j} (u_j^\top k(Z, z_i))^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} \sum_{j=s+1}^{\ell} \lambda_j (u_i^\top \mathbf{1}_j)^2 = \frac{1}{\ell} \sum_{j=s+1}^{\ell} \lambda_j, \tag{5.20}$$

and the proof is complete. □

5.B.4 Proof of Lemma 5.3

Proof. From the min-max principle, we have

$$\lambda_j = \min_{\substack{V_{j-1} \subset \mathbb{R}^\ell \\ \dim V_{j-1} \le j-1}} \max_{x_j \in V_{j-1}^{\perp},\ \|x_j\|_2 = 1} x_j^\top k(Z, Z) x_j, \tag{5.21}$$

where V_{j-1} is a linear subspace of $\mathbb{R}^\ell$. Recall the Mercer expansion $k(x, y) = \sum_{i=1}^{\infty} \sigma_i e_i(x) e_i(y)$. By letting $e_j(Z) = (e_j(z_1), \dots, e_j(z_\ell))^\top \in \mathbb{R}^\ell$, we can write

$k(Z,Z) = \sum_{i=1}^\infty \sigma_i e_i(Z)e_i(Z)^\top$. We assume that this equality holds in the following. We especially write the remainder term as $k_{s+1}(Z,Z) := k(Z,Z) - \sum_{i=1}^s \sigma_i e_i(Z)e_i(Z)^\top$

Consider taking $V_s = \operatorname{span}\{e_1(Z), \ldots, e_s(Z)\}$ and

$$x_j \in \operatorname*{argmax}_{x \in V_{j-1}^\perp,\ \|x\|_2=1} x^\top k(Z,Z)x, \qquad V_j = \operatorname{span}(V_{j-1} \cup \{x_j\})$$

for $j = s+1, \ldots, \ell$ in (5.21). Then, $\lambda_j' := x_j^\top k(Z,Z)x$ satisfies $\lambda_j \le \lambda_j'$, and so we have

$$\sum_{j=s+1}^\ell \lambda_j \le \sum_{j=s+1}^\ell \lambda_k' = \sum_{j=s+1}^\ell x_j^\top k(Z,Z)x_j = \sum_{j=s+1}^\ell x_j^\top k_{s+1}(Z,Z)x_j,$$

where we have used that $x_j^\top e_i(Z) = 0$ for any $i \le s < j$ in the last inequality. By taking some $\{x_1, \ldots, x_s\} \subset \mathbb{R}^\ell$, we can make $\{x_1, \ldots, x_\ell\}$ a orthonormal basis of $\mathbb{R}^\ell$, so we obtain

$$\sum_{j=s+1}^\ell \lambda_j \le \sum_{j=s+1}^\ell x_j^\top k_{s+1}(Z,Z)x_j \le \sum_{j=1}^\ell x_j^\top k_{s+1}(Z,Z)x_j = \operatorname{tr} k_{s+1}(Z,Z).$$

Therefore, we have

$$\frac{1}{\ell}\sum_{j=s+1}^\ell \lambda_j \le \frac{1}{\ell} \operatorname{tr} k_{s+1}(Z,Z) = \frac{1}{\ell}\sum_{i=1}^\ell k_{s+1}(z_i, z_i),$$

and we obtain the desired inequality in expectation since $\mathbb{E}[k_{s+1}(z_i,z_i)] = \sum_{j=s+1}^\infty \sigma_j$. □

5.B.5 Proof of Theorem 5.4

We first prove the following generic proposition by exploiting the ingredients given in Section 5.A.

Proposition 5.20. *Let Q be an arbitrary deterministic m-dimensional orthogonal projection in $\mathcal{H}_k$. Then, for any random orthogonal projection P possibly depending on Z, we have*

$$\mu(\|PQk(\cdot,x)\|_{\mathcal{H}_k}) \le \mu_Z(\|PQk(\cdot,x)\|_{\mathcal{H}_k}) + \sqrt{\frac{k_{\max}}{\ell}}\left(36m + \sqrt{2\log\frac{1}{\delta}}\right) \tag{5.22}$$

with probability at least $1-\delta$.

Furthermore, with regard to the expectation, we also have

$$\mathbb{E}[\mu(\|PQk(\cdot,x)\|_{\mathcal{H}_k})] \le 2\mathbb{E}[\mu_Z(\|PQk(\cdot,x)\|_{\mathcal{H}_k})] + \frac{\sqrt{k_{\max}}}{\ell}\left(\frac{80m^2\log(1+2\ell)}{9}+69\right). \tag{5.23}$$

Proof. Let $\{v_1,\ldots,v_m\}$ be an orthonormal basis of $Q\mathcal{H}_k$. Let also $\{u_i\}_{i\in I}$ and $\{u_i\}_{i\in J}$ be respectively an orthonormal basis of $P\mathcal{H}_k$ and $(P\mathcal{H}_k)^\perp$, so $\{u_i\}_{i\in I\cup J}$ is an orthonormal basis of $\mathcal{H}_k$.

Let us compute $\|PQk(\cdot,x)\|_{\mathcal{H}_k}^2$. Since we have

$$PQk(\cdot,x) = P\left(\sum_{j=1}^m \langle v_j, k(\cdot,x)\rangle_{\mathcal{H}_k} v_j\right) = \sum_{j=1}^m v_j(x)Pv_j = \sum_{i\in I}\sum_{j=1}^m v_j(x)\langle u_i, v_j\rangle_{\mathcal{H}_k} u_i$$

(where we can exchange the summation as they converge in $\mathcal{H}_k$), we obtain

$$\|PQk(\cdot,x)\|_{\mathcal{H}_k}^2 = \sum_{i\in I}\left(\sum_{j=1}^m v_j(x)\langle u_i, v_j\rangle_{\mathcal{H}_k}\right)^2 = \|A_{P,Q}\boldsymbol{v}_x\|_{\ell^2(I)}^2 = \boldsymbol{v}_x^\top A_{P,Q}^* A_{P,Q}\boldsymbol{v}_x,$$

where $\boldsymbol{v}_x = (v_1(x),\ldots,v_m(x))^\top \in \mathbb{R}^m$ and $A_{P,Q}$ is a linear operator $\mathbb{R}^m \to \ell^2(I)$ given by $a=(a_1,\ldots,a_m)^\top \mapsto (\sum_{j=1}^m \langle u_i, v_j\rangle_{\mathcal{H}_k} a_j)_{i\in I}$, and $A_{P,Q}^* : \ell^2(I) \to \mathbb{R}^m$ is its dual (defined by the property $\langle a, A_{P,Q}^* b\rangle_{\mathbb{R}^m} = \langle A_{P,Q}a, b\rangle_{\ell^2(I)}$), which can be understood as the "transpose" of $A_{P,Q}$. Note that $A_{P,Q}^*A_{P,Q}$ can be regarded as an $m\times m$ matrix and we have

$$(A_{P,Q}^*A_{P,Q})_{j,h} = \sum_{i\in I}\langle u_i, v_j\rangle_{\mathcal{H}_k}\langle u_i, v_h\rangle_{\mathcal{H}_k} = \langle Pv_j, Pv_h\rangle_{\mathcal{H}_k}.$$

We can also define $B_{P,Q} = A_{P^\perp,Q}$ by replacing P with $P^\perp$. Then we have

$$(A_{P,Q}^*A_{P,Q})_{j,h} + (B_{P,Q}^*B_{P,Q})_{j,h} = \langle Pv_j, Pv_h\rangle_{\mathcal{H}_k} + \langle P^\perp v_j, P^\perp v_h\rangle_{\mathcal{H}_k} = \langle v_j, v_h\rangle_{\mathcal{H}_k} = \delta_{jh},$$

so $A_{P,Q}^\top A_{P,Q}$ is an $m\times m$ positive semi-definite matrix with $A_{P,Q}^\top A_{P,Q} \le I_m$.

It thus suffices to consider a uniform estimate of $\mu(\sqrt{\boldsymbol{v}_x^\top S\boldsymbol{v}_x}) - \mu_Z(\sqrt{\boldsymbol{v}_x^\top S\boldsymbol{v}_x})$ with a positive semi-definite matrix $S \le I_m$. This S can be written as $S = U^\top U$ by using a $U\in\mathbb{R}^{m\times m}$ with $\|U\|_2 \le 1$, so we shall solve the following problem:

> Find a uniform upper bound of $\mu(\|U\boldsymbol{v}_x\|_2)-\mu_Z(\|U\boldsymbol{v}_x\|_2)$ for any matrix $U \in \mathbb{R}^{m\times m}$ with $\|U\|_2 \le 1$.

Now we can reduce our problem to a routine work of bounding the covering number of the function class $\mathcal{F} := \{f_U := x \mapsto \|U\boldsymbol{v}_x\|_2 \mid U \in \mathcal{U}\}$, where $\mathcal{U} := \{U \in \mathbb{R}^{m\times m} \mid \|U\|_2 \le 1\}$.

For any $x \in \mathcal{X}$, we have

$$\|\boldsymbol{v}_x\|_2^2 = \sum_{j=1}^{\ell} v_j(x)^2 = \|Qk(\cdot,x)\|_{\mathcal{H}_k}^2 \le \|k(\cdot,x)\|_{\mathcal{H}_k}^2 = k(x,x).$$

If $\mathcal{U}_\delta$ is a δ-covering of $\mathcal{U}$, then $\{f_U\}_{U\in\mathcal{U}_\delta}$ gives a $\delta\sqrt{k_{\max}}$-covering. Indeed, for any $U, V \in \mathcal{U}$ with $\|U-V\|_2 \le \delta$, we have

$$\begin{aligned} d_Z(f_U,f_V)^2 &= \frac{1}{\ell}\sum_{i=1}^{\ell}(\|U\boldsymbol{v}_{z_i}\|_2 - \|V\boldsymbol{v}_{z_i}\|_2)^2 \\ &\le \frac{1}{\ell}\sum_{i=1}^{\ell}\|(U-V)\boldsymbol{v}_{z_i}\|_2^2 \le \delta^2\frac{1}{\ell}\sum_{i=1}^{\ell}\|\boldsymbol{v}_{z_i}\|_2^2 \le \delta^2 k_{\max}. \end{aligned}$$

Here, we have the covering number bound $\log\mathcal{N}(\mathcal{U},\|\cdot\|_2;\delta) \le m^2\log\left(1+\frac{2}{\delta}\right)$ for $\delta \le 1$ (and 0 for $\delta \ge 1$) as $\mathcal{U}$ can be seen as a unit ball of $\mathbb{R}^{m^2}$ in a certain norm [174, Example 5.8], so $\log\mathcal{N}(\mathcal{F},d_Z;\varepsilon) \le m^2\log(1+2\sqrt{k_{\max}}/\varepsilon)$ for $\varepsilon \le \sqrt{k_{\max}}$.

Therefore, from Proposition 5.18, we have

$$\begin{aligned} \mathcal{R}_Z(\mathcal{F}) &\le \frac{12}{\sqrt{\ell}}\int_0^{\sqrt{k_{\max}}}\sqrt{m^2\log\left(1+\frac{2\sqrt{k_{\max}}}{\varepsilon}\right)}\,\mathrm{d}\varepsilon \\ &= \frac{12m\sqrt{k_{\max}}}{\sqrt{\ell}}\int_0^1\sqrt{\log\left(1+\frac{2}{t}\right)}\,\mathrm{d}t \le \frac{18m\sqrt{k_{\max}}}{\sqrt{\ell}}, \end{aligned}$$

where we have used the estimate

$$\int_0^1\sqrt{\log\left(1+\frac{2}{t}\right)}\,\mathrm{d}t \le \int_0^1\frac{1}{2}\left(1+\log\left(1+\frac{2}{t}\right)\right)\mathrm{d}t = \frac{1}{2}+\frac{1}{2}\log\frac{27}{4} \le \frac{3}{2}.$$

Since we also have a bound $\|f_U\|_\infty \le \|U\|_2\sqrt{k_{\max}}$, we can use Proposition 5.17

to obtain

$$\begin{aligned}\mu(\|PQk(\cdot,x)\|_{\mathcal{H}_k}) - \mu_Z(\|PQk(\cdot,x)\|_{\mathcal{H}_k}) &\le \sup_{f\in\mathcal{F}}(\mu_Z(f)-\mu(f)) \\ &\le \sqrt{\frac{k_{\max}}{\ell}}\left(36m + \sqrt{2\log\frac{1}{\delta}}\right)\end{aligned}$$

with probability at least $1-\delta$. So we have proven (5.22).

We next prove (5.23) by using Proposition 5.19. We have the same bound for $\log\mathcal{N}(\mathcal{F}, \|\cdot\|_{L^1}(\mu);\varepsilon)$ from the same argument as above, and so we especially get

$$\log N\left(\mathcal{F}, \|\cdot\|_{L^1(\mu)}; \frac{\sqrt{k_{\max}}}{\ell}\right) \le m^2\log(1+2\ell).$$

As $\|f\|_{L^\infty(\mu)} \le \sqrt{k_{\max}} =: F$ holds for all $f\in\mathcal{F}$, we can now apply Proposition 5.19 with $\varepsilon = F/\ell$ to obtain the desired conclusion. □

We next prove the following proposition that includes the desired assertion by using Proposition 5.20.

Proposition 5.21. *Let $Z=(z_i)_{i=1}^\ell$ be an ℓ-point independent sample from μ. Let P be a random orthogonal projection in $\mathcal{H}_k$ possibly depending on Z. For any integer $m\ge 1$, with probability at least $1-\delta$, we have*

$$\int_{\mathcal{X}}\|Pk(\cdot,x)\|_{\mathcal{H}_k}\,\mathrm{d}\mu(x) \le \frac{1}{\ell}\sum_{i=1}^{\ell}\|Pk(\cdot,z_i)\|_{\mathcal{H}_k} + \sqrt{\frac{k_{\max}}{\ell}}\left(36m+\sqrt{\frac{9}{2}\log\frac{2}{\delta}}\right) + 3\sqrt{\sum_{j>m}\sigma_j}.$$

Furthermore, in expectation, we have the following bound:

$$\begin{aligned}\mathbb{E}\left[\int_{\mathcal{X}}\|Pk(\cdot,x)\|_{\mathcal{H}_k}\,\mathrm{d}\mu(x)\right] &\le \mathbb{E}\left[\frac{2}{\ell}\sum_{i=1}^{\ell}\|Pk(\cdot,z_i)\|_{\mathcal{H}_k}\right] \\ &\quad + \frac{\sqrt{k_{\max}}}{\ell}\left(\frac{80m^2\log(1+2\ell)}{9}+69\right) + 4\sqrt{\sum_{j>m}\sigma_j}. \end{aligned} \tag{5.24}$$

Proof. Note that we use the fact that for any projection operator P $\|Pf\|\le\|f\|$ frequently within the proof. For an ℓ-point sample $Z=(z_1,\dots,z_\ell)\subset\mathcal{X}$, let us denote μ_Z be the mapping $f \mapsto \frac{1}{\ell}\sum_{i=1}^{\ell} f(z_i)$. If we have $f_-, f \in L^1(\mu)$ with $f_- \le f$, we can generally obtain

$$\begin{aligned} \mu(f) - \mu_Z(f) &= (\mu(f) - \mu(f_-)) + (\mu(f_-) - \mu_Z(f_-)) + (\mu_Z(f_-) - \mu_Z(f)) \\ &\le \mu(f - f_-) + (\mu(f_-) - \mu_Z(f_-)). \end{aligned} \tag{5.25}$$

We here use $f(x) = \|Pk(\cdot, x)\|_{\mathcal{H}_k}$ and $f_-(x) = \|PP_m k(\cdot, x)\|_{\mathcal{H}_k} - \|PP_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}$ for an m, where P_m is the projection operator onto $\operatorname{span}\{e_1, \dots, e_m\}$ in $\mathcal{H}_k$ and $P_m^\perp$ is its orthogonal complement. In this case, $\mu(f - f_-)$ can easily be estimated by Cauchy–Schwarz as follows:

$$\begin{aligned} \mu(f - f_-) &\le \mu(2\|PP_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}) \le 2\mu(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}) \\ &\le 2\sqrt{\mu(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}^2)} = 2\sqrt{\sum_{j>m} \sigma_j}, \end{aligned} \tag{5.26}$$

where we have used the fact

$$\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}^2 = \|k(\cdot, x)\|_{\mathcal{H}_k}^2 - \|P_m k(\cdot, x)\|_{\mathcal{H}_k}^2 = k(x, x) - \sum_{i=1}^{m} \sigma_i e_i(x)^2 = \sum_{i=m+1}^{\infty} \sigma_i e_i(x)^2.$$

We also bound $\mu(f_-) - \mu_Z(f_-)$ by

$$\mu(f_-) - \mu_Z(f_-) \le \mu(\|PP_m k(\cdot, x)\|_{\mathcal{H}_k}) - \mu_Z(\|PP_m k(\cdot, x)\|_{\mathcal{H}_k}) + \mu_Z(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}), \tag{5.27}$$

where we have used the second inequality in (5.26) for μ_Z. The last term $\mu_Z(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k})$ above is estimated either in expectation or in high probability as follows:

$$\begin{cases} \mathbb{E}\left[\mu_Z(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k})\right] \le \sqrt{\displaystyle\sum_{j>m} \sigma_j}. \\ \mu_Z(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}) \le \sqrt{\displaystyle\sum_{j>m} \sigma_j} + \sqrt{\dfrac{k_{\max}}{2\ell} \log \dfrac{1}{\delta}} \text{ with probability at least } 1 - \delta. \end{cases} \tag{5.28}$$

The latter follows from a simple calculation of Hoeffing's inequality.

Thus, it suffices to derive a bound for $\mu(\|PP_m k(\cdot, x)\|_{\mathcal{H}_k}) - \mu_Z(\|PP_m k(\cdot, x)\|_{\mathcal{H}_k})$ or its expectation; we do it by letting $Q = P_m$ and $\hat{f} = f$ in Proposition 5.20. By combining (just summing up) the inequalities (5.25)–(5.28), and (5.22), we obtain the desired inequality in high probability. For the result in expectation, we first combine the inequalities (5.25)–(5.28), and (5.23) to get the bound

$$\mathbb{E}[\mu(f)]-\mathbb{E}[\mu_Z(f)] \leq \mathbb{E}[\mu_Z(\|PP_m k(\cdot,x)\|_{\mathcal{H}_k})]+\frac{\sqrt{k_{\max}}}{\ell}\left(\frac{80m^2\log(1+2\ell)}{9}+69\right)+3\sqrt{\sum_{j>m}\sigma_j}$$

(recall $f(x)=\|Pk(\cdot,x)\|_{\mathcal{H}_k}$). Since we can also estimate $\mathbb{E}[\mu_Z(\|PP_m k(\cdot,x)\|_{\mathcal{H}_k})]$ as

$$\begin{aligned}\mathbb{E}[\mu_Z(\|PP_m k(\cdot,x)\|_{\mathcal{H}_k})] &\leq \mathbb{E}[\mu_Z(\|Pk(\cdot,x)\|_{\mathcal{H}_k})]+\mathbb{E}\big[\mu_Z(\|PP_m^\perp k(\cdot,x)\|_{\mathcal{H}_k})\big]\\ &\leq \mathbb{E}[\mu_Z(\|Pk(\cdot,x)\|_{\mathcal{H}_k})]+\sqrt{\sum_{j>m}\sigma_j},\end{aligned}$$

we obtain the desired conclusion. □

5.B.6 Proof of Remark 5.1

Proof. We assume $\ell \geq 3$ here. Let $F(x) := -\beta^{-1}x^{1-1/d}\exp(-\beta x^{1/d})$. If $d\geq 2$, its derivative is

$$F'(x)=\exp(-\beta x^{1/d})-\frac{1-1/d}{\beta}x^{-1/d}\exp(-\beta x^{1/d})=\left(1-\frac{1-1/d}{\beta}x^{-1/d}\right)\exp(-\beta x^{1/d}).$$

Thus, if $x\geq(\log\ell)^d/\beta^d$, we have $F'(x)\geq d\exp(-\beta x^{1/d})$. This inequality is still true if $d=1$. By taking $m=\lfloor(2\log\ell)^d/\beta^d\rfloor$, we obtain

$$\sum_{i>m}\sigma_i \lesssim \int_{2(\log\ell)^d/\beta^d}^{\infty}\exp(-\beta x^{1/d})\,\mathrm{d}x \leq -dF(2(\log\ell)^d/\beta^d)=\frac{2^{d-1}d}{\beta^d}\cdot\frac{(\log\ell)^{d-1}}{\ell^2}.$$

Therefore, this choice of m satisfies

$$\sqrt{\sum_{i>m}\sigma_i}=\mathcal{O}\left(\frac{(\log\ell)^{(d-1)/2}}{\ell}\right), \qquad m^2=\mathcal{O}\big((\log\ell)^{2d}\big).$$

Combining these with the inequality in Corollary 5.5 gives the desired estimate. □

5.B.7 Proof of Proposition 5.6

Proof. We basically just compute the trace of the operator $P_Z^\perp\mathcal{K}$. Indeed, we have

$$\int_{\mathcal{X}}\|P_Z^\perp k(\cdot,x)\|_{\mathcal{H}_k}^2=\int_{\mathcal{X}}(k(x,x)-k^Z(x,x))\,\mathrm{d}\mu(x), \tag{5.29}$$

and, from (5.5), we also have the following identity:

$$\int_{\mathcal{X}} k(x,x)\,\mathrm{d}\mu(x) = \sum_{i=1}^{\infty} \langle e_i, \mathcal{K}e_i\rangle_{L^2(\mu)}. \tag{5.30}$$

For k^Z, as we can write $k^Z(x,y) = \sum_{i=1}^{\ell} g_i(x)g_i(y)$ by using $g_i \in L^2(\mu)$ (see e.g., (5.18)), we can also have

$$\int_{\mathcal{X}} k^Z(x,x)\,\mathrm{d}\mu(x) = \sum_{i\in I} \left\langle e_i, \mathcal{K}^Z e_i\right\rangle_{L^2(\mu)} = \sum_{i=1}^{\infty} \left\langle e_i, \mathcal{K}^Z e_i\right\rangle_{L^2(\mu)}, \tag{5.31}$$

where $\mathcal{K}^Z : L^2(\mu) \to L^2(\mu)$ is the integral operator given by $g \mapsto \int_{\mathcal{X}} k^Z(\cdot,x)g(x)\,\mathrm{d}\mu(x)$, and $(e_i)_{i\in I}$ is an orthonormal basis of $L^2(\mu)$ including $(e_i)_{i=1}^{\infty}$. The second equality follows from the fact that $\mathcal{K}-\mathcal{K}^Z$ is a (semi-)positive definite operator since $k-k^Z$ is a positive definite kernel, and so we have $0 \le \left\langle e_i, \mathcal{K}^Z e_i\right\rangle_{L^2(\mu)} \le \langle e_i, \mathcal{K}e_i\rangle_{L^2(\mu)} = 0$ for any $i \in I \setminus \mathbb{Z}_{>0}$. For this integral operator, since we have $k^Z(\cdot,x) = P_Z k(\cdot,x)$, we can prove

$$\mathcal{K}^Z g = \int_{\mathcal{X}} P_Z k(\cdot,x)g(x)\,\mathrm{d}\mu(x) = P_Z \int_{\mathcal{X}} k(\cdot,x)g(x)\,\mathrm{d}\mu(x) = P_Z\mathcal{K}g$$

for any $g \in L^2(\mu)$ under the well-definedness of $\mathcal{K}$. Thus, from (5.29)–(5.31), we have

$$\int_{\mathcal{X}} \|P_Z^{\perp} k(\cdot,x)\|_{\mathcal{H}_k}^2 = \sum_{i=1}^{\infty} \left\langle e_i, (\mathcal{K}-\mathcal{K}^Z)e_i\right\rangle_{L^2(\mu)} = \sum_{i=1}^{\infty} \left\langle e_i, P_Z^{\perp}\mathcal{K}e_i\right\rangle_{L^2(\mu)}. \tag{5.32}$$

For general $f \in \mathcal{H}_k$ and $g \in L^2(\mu)$, we can prove

$$\langle f, \mathcal{K}g\rangle_{\mathcal{H}_k} = \left\langle f, \int_{\mathcal{X}} k(\cdot,x)g(x)\,\mathrm{d}\mu(x)\right\rangle_{\mathcal{H}_k} = \int_{\mathcal{X}} \langle f, k(\cdot,x)\rangle_{\mathcal{H}_k} g(x)\,\mathrm{d}\mu(x) = \langle f, g\rangle_{L^2(\mu)},$$

so that in particular

$$\left\langle g, P_Z^{\perp}\mathcal{K}g\right\rangle_{L^2(\mu)} = \left\langle \mathcal{K}g, P_Z^{\perp}\mathcal{K}g\right\rangle_{\mathcal{H}_k} = \|P_Z^{\perp}\mathcal{K}g\|_{\mathcal{H}_k}^2.$$

By letting $g = e_i$ in the above equation, we can deduce the desired equality from (5.32). For the inequality, use the bound

$$\|P_Z^{\perp}\mathcal{K}e_i\|_{\mathcal{H}_k}^2 \le \|\mathcal{K}e_i\|_{\mathcal{H}_k}^2 = \|\sigma_i e_i\|_{\mathcal{H}_k}^2 = \sigma_i\|\sqrt{\sigma_i}e_i\|_{\mathcal{H}_k}^2 = \sigma_i$$

for each $i > m$. □

5.B.8 Proof of Corollary 5.7

Proof. From Proposition 5.6 and (5.8), it suffices to prove for an arbitrary $g \in L^2(\mu)$ that

$$\|P_Z^\perp \mathcal{K} g\|_{\mathcal{H}_k}^2 = \inf_{w_i} \sup_{\|f\|_{\mathcal{H}_k} \le 1} \left| \mu(fg) - \sum_{i=1}^{\ell} w_i f(z_i) \right|^2 \le 4 \sum_{i>\ell} \sigma_i.$$

It is indeed an immediate consequence of Belhadji [15, Theorem 4]. □

5.B.9 Proof of Lemma 5.8

Proof. Given the Mercer decomposition $k(x,y) = \sum_{i=1}^\infty \sigma_i e_i(x) e_i(y)$, we can compute

$$\begin{aligned} h_\mu(x,y) &= \int_{\mathcal{X}} k(x,t)k(t,y) \,\mathrm{d}\mu(t) \\ &= \sum_{i,j=1}^\infty \sigma_i \sigma_j e_i(x) e_j(y) \int_{\mathcal{X}} e_i(t) e_i(t) \,\mathrm{d}\mu(t) \\ &= \sum_{i,j=1}^\infty \delta_{ij} \sigma_i \sigma_j e_i(x) e_j(y) = \sum_{i=1}^\infty \sigma_i^2 e_i(x) e_i(y), \end{aligned}$$

where we have used the fact that $(e_i)_{i=1}^\infty$ is an orthonormal set in $L^2(\mu)$. □

5.B.10 Proof of Lemma 5.9

Proof. From (5.9), we have

$$\langle f_i, f_j \rangle_{L^2(\mu)} = v_i^\top (H^+)^\top H^\top H H^+ v_j = (HH^+ v_i)^\top (HH^+ v_j). \tag{5.33}$$

Here, note that $\{v_i,\, \kappa_i > 0\} \subset (\ker H^\top)^\perp$ as we have, for any $v \in \ker H^\top$,

$$0 = v^\top H k(Z,Z)^+ H^\top v = \sum_{i=1}^{\ell} \kappa_i v^\top v_i v_i^\top v = \sum_{i=1}^{\ell} \kappa_i (v^\top v_i)^2.$$

Therefore, $HH^+ v_i = v_i$ if $\kappa_i > 0$ since HH^+ is the projection onto $(\ker H^\top)^\perp$, and so $\{f_i,\, \kappa_i > 0\}$ is orthonormal from (5.33). We can also see that $f_i = (H^+ v_i)^\top k(Z, \cdot)$ is an eigenfunction of $\mathcal{K}^Z$ from the remark below (5.10) and $HH^+ v_i = v_i$. □

5.B.11 Proof of Proposition 5.10

Proof. We rewrite k_μ^Z in terms of another summation as follows:

$$\begin{aligned}
k_\mu^Z(x,y) &:= \sum_{i=1}^{\ell} \kappa_i f_i(x) f_i(y) \\
&= k(x,Z)H^+ \left(\sum_{i=1}^{\ell} \kappa_i v_i v_i^\top \right) (H^\top)^+ k(Z,y) \\
&= k(x,Z)H^+ H k(Z,Z)^+ H^\top (H^\top)^+ k(Z,y) \\
&= \sum_{\lambda_i > 0} \frac{1}{\lambda_i} u_i^\top H^\top (H^+)^\top k(Z,x) k(y,Z) H^+ H u_i, \qquad (5.34)
\end{aligned}$$

where (λ_i, u_i) are eigenpairs of $k(Z,Z)$. Recall also that we have

$$k^Z(x,y) = k(x,Z)k(Z,Z)^+ k(Z,y) = \sum_{\lambda_i>0} \frac{1}{\lambda_i} u_i^\top k(Z,x) k(y,Z) u_i. \qquad (5.35)$$

From (5.34) and this, it suffices to prove $u^\top k(Z,\cdot) = u^\top H^\top (H^+)^\top k(Z,\cdot)$ in $L^2(\mu)$ for any $u \in \mathbb{R}^\ell$. Indeed, we have

$$\begin{aligned}
&\int_{\mathcal{X}} \left(u^\top k(Z,x) - u^\top H^\top (H^+)^\top k(Z,x) \right)^2 \mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \left(u^\top \left(I_\ell - H^\top (H^+)^\top \right) k(Z,x) \right)^2 \mathrm{d}\mu(x) \\
&= u^\top \left(I_\ell - H^\top (H^+)^\top \right) \left(\int_{\mathcal{X}} k(Z,x) k(x,Z) \, \mathrm{d}\mu(x) \right) (I_\ell - H^+ H) u \\
&= u^\top \left(I_\ell - H^\top (H^+)^\top \right) H^\top H (I_\ell - H^+ H) u = 0
\end{aligned}$$

since $H^\top (H^+)^\top H^\top = H^\top$ and $HH^+H = H$ hold (I_ℓ is the identity matrix). Thus, we obtain the desired assertion.

Finally, we prove that k_μ^Z and k^Z coincide when $\ker h_\mu(Z,Z) \subset \ker k(Z,Z)$. From (5.34) and (5.35), it suffices to prove $H^+ H u_i = u_i$ for indices i with $\lambda_i > 0$. Note that H^+H is the orthogonal projection onto the orthogonal complement of $\ker H = \ker H^\top H = h_\mu(Z,Z)$ from a general property of the pseudo-inverse. Since u_i is an eigenvector of $k(Z,Z)$ with a positive eigenvalue λ_i, it is orthogonal to any $v \in \ker k(Z,Z)$ (as $u_i^\top v = \lambda_i^{-1} u_i^\top k(Z,Z) v = 0$). Therefore, if we have $\ker h_\mu(Z,Z) \subset \ker k(Z,Z)$, u_i is also orthogonal to $\ker h_\mu(Z,Z)$ and so $H^+ H u_i = u_i$ as desired. □

5.B.12 Proof of Proposition 5.13

First, we give a proof for a folklore property of products of positive semi-definite matrices.

Lemma 5.22. *Let $\ell, m \geq n$ be positive integers and $A, B \in \mathbb{R}^{n\times n}$ be (symmetric) positive semi-definite matrices. Assume $B = C^\top C = D^\top D$ for a real matrix $C \in \mathbb{R}^{m\times n}$ and $D \in \mathbb{R}^{\ell\times n}$. Then, $CAC^\top$ and $DAD^\top$ have the same set of nonzero eigenvalues with the same multiplicity (in terms of real eigenvectors).*

Proof. For a real square matrix $M \in \mathbb{R}^{j\times j}$ and a real number λ, let us define $S_\lambda(M) := \{v \in \mathbb{R}^j \mid Mv = \lambda v\}$ be the real eigenspace of M corresponding to λ.

We shall prove there is a bijection between $S_\lambda(AB)$ and $S_\lambda(CAC^\top)$ for each real $\lambda \neq 0$ (and the same for $S_\lambda(DAD^\top)$ by symmetry). Once we establish this, we see that each $\lambda \neq 0$ has the same multiplicity as an eigenvalue of $CAC^\top$ and $DAD^\top$ (multiplicity can be zero; in that case λ is not an eigenvalue), and the desired assertion follows.

Let us fix $\lambda \neq 0$. If $v \in S_\lambda(CAC^\top)$, we have $CAC^\top(Cv) = CABv = \lambda(Cv)$, so $Cv \in S_\lambda(CAC^\top)$. We also have $Cv' \neq Cv$ for another element $(v \neq)v' \in S_\lambda(AB)$ since $AC^\top(Cv' - Cv) = AB(v' - v) = \lambda(v' - v) \neq 0$. Thus, matrix multiplication by C is an injective map from $S_\lambda(AB)$ to $S_\lambda(CAC^\top)$.

Let us finally prove $S_\lambda(AB) \ni v \mapsto Cv \in S_\lambda(CAC^\top)$ is surjective. Let $u \in S_\lambda(CAC^\top)$. Then, $u = \lambda^{-1}(\lambda u) = \lambda^{-1}CAC^\top u = C(\lambda^{-1}AC^\top u)$, so we can write $u = Cv$ for $v = \lambda^{-1}AC^\top u$. It remains to prove $v \in S_\lambda(AB)$, but we can see it as follows:

$$ABv = AB\left(\frac{1}{\lambda}AC^\top u\right) = \frac{1}{\lambda}(AC^\top C)AC^\top u = \frac{1}{\lambda}AC^\top(CAC^\top u) = \frac{1}{\lambda}AC^\top(\lambda u) = \lambda v.$$

Therefore, we have a bijection between $S_\lambda(AB)$ and $S_\lambda(CAC^\top)$ and we are done. □

Recall $\mu(k_\mu^Z - k_{s,\mu}^Z) \leq \sum_{i=s+1}^\ell \kappa_i$ holds for eigenvalues $\kappa_1 \geq \cdots \kappa_\ell \geq 0$ of $H_\mu k(Z,Z)^+ H_\mu^\top$ with $H_\mu^\top H_\mu = h_\mu(Z,Z)$ (that immediately follows from the definitions of k_μ^Z and $k_{s,\mu}^Z$, and that f_i are $L^2(\mu)$-orthonormal). By replacing μ with μ_X, we have $\mu_X(k_X^Z - k_{s,X}^Z) \leq \sum_{i=s+1}^\ell \kappa_i^X$ for eigenvalues of $\kappa_1^X \geq \cdots \geq \kappa_\ell^X \geq 0$ of $H_X k(Z,Z)^+ H_X^\top$, where $H_X^\top H_X = h_X(Z,Z) = \frac{1}{M}k(Z,X)k(X,Z)$.

By using the lemma, we can see that κ_i^X are actually the same as the eigenvalues of $\frac{1}{M}k(X,Z)k(Z,Z)^+k(Z,X) = \frac{1}{M}k^Z(X,X)$. As $k-k^Z$ is a positive definite kernel, $k(X,X)-k^Z(X,X)$ is a positive semi-definite matrix, the i-th largest eigenvalue of $k^Z(X,X)$ is bounded by the i-th largest eigenvalue of $k(X,X)$ (Weyl's inequality).

Now, let $\lambda_1^X \ge \lambda_2^X \ge \cdots \ge 0$ be the eigenvalues of $k(X,X)$. From the above argument, we have

$$\mu_X(k_X^Z - k_{s,X}^Z) \le \sum_{i=s+1}^{\ell} \kappa_i^X \le \frac{1}{M}\sum_{i=s+1}^{\ell} \lambda_i^X \le \frac{1}{M}\sum_{i=s+1}^{M} \lambda_i^X.$$

Notice that we can apply Lemma 5.3 with X instead of Z, and obtain $\mathbb{E}\bigl[\mu_X(k_X^Z - k_{s,X}^Z)\bigr] \le \sum_{i>s}\sigma_i$ as desired.

5.B.13 Proof of Proposition 5.14

Proof. Fix a sample X with $\ker k(X,Z) \subset \ker k(Z,Z)$ and let us use the same notation as in μ, i.e.,

- $H^\top H = h_X(Z,Z) = \frac{1}{M}k(Z,X)k(X,Z)$;
- $Hk(Z,Z)^+H^\top = V\operatorname{diag}(\kappa_1,\dots,\kappa_\ell)V^\top$ with $\kappa_1 \ge \cdots \kappa_\ell \ge 0$ and V being orthogonal;
- $f_i = (H^+v_i)^\top k(Z,\cdot)$ and $k_X^Z(x,y) = \sum_{i=1}^{\ell}\kappa_i f_i(x)f_i(y)$.

In this case, from the same argument as the last paragraph in the proof of Proposition 5.10, we have H^+H is an identity map over $(\ker h_X(Z,Z))^\perp = (\ker k(X,Z))^\perp \supset (\ker k(Z,Z))^\perp$. By considering the SVD of $k(Z,Z)$, we see that $(\ker k(Z,Z))^\perp$ is exactly the linear subspace of $\mathbb{R}^\ell$ spanned by eigenvectors of $k(Z,Z)$ with nonzero eigenvalues, which is equal to $\{k(Z,Z)v \mid v \in \mathbb{R}^\ell\} = \{k(Z,Z)^+v \mid v \in \mathbb{R}^\ell\}$. In particular, we have $H^+Hk(Z,Z)^+ = k(Z,Z)^+$.

We now prove that $\{\sqrt{\kappa_i}f_i \mid i \ge 1,\ \kappa_i > 0\}$ actually forms an orthonoramal set in $\mathcal{H}_k$. Indeed, if $\kappa_i, \kappa_j > 0$, we have

$$\begin{aligned}
\big\langle \sqrt{\kappa_i} f_i, \sqrt{\kappa_j} f_j \big\rangle_{\mathcal{H}_k} &= \sqrt{\kappa_i \kappa_j} v_i^\top (H^+)^\top k(Z,Z) H^+ v_j \\
&= \frac{1}{\sqrt{\kappa_i \kappa_j}} v_i^\top \left[H k(Z,Z)^+ H^\top \right] (H^+)^\top k(Z,Z) H^+ \left[H k(Z,Z)^+ H^\top \right] v_j \\
&= \frac{1}{\sqrt{\kappa_i \kappa_j}} v_i^\top H k(Z,Z)^+ k(Z,Z) k(Z,Z)^+ H^\top v_j \\
&= \frac{1}{\sqrt{\kappa_i \kappa_j}} v_i^\top H k(Z,Z)^+ H^\top v_j = \delta_{ij},
\end{aligned}$$

where we have used the fact that v_i and v_j are eigenvectors of $Hk(Z,Z)^+H^\top$ with eigenvalues κ_i and κ_j, respectively.

Let $P : \mathcal{H}_k \to \mathcal{H}_k$ be the orthogonal projection onto $\operatorname{span}\{\sqrt{\kappa_i} f_i \mid i > s,\ \kappa_i > 0\}$. Then, we have

$$Pk(\cdot, x) = \sum_{i=s+1}^{\ell} \langle \sqrt{\kappa_i} f_i, k(\cdot, x) \rangle_{\mathcal{H}_k} \sqrt{\kappa_i} f_i = \sum_{i=s+1}^{\ell} \sqrt{\kappa_i} f_i(x) \sqrt{\kappa_i} f_i,$$

and so $\|Pk(\cdot,x)\|_{\mathcal{H}_k}^2 = \sum_{i=s+1}^{\ell} \kappa_i f_i(x)^2 = k^Z(x,x) - k^Z_{s,X}(x,x)$. Note that the projection P is a random operator depending on the sample X. Now, we can use Theorem 5.4 with the empirical measure given by X instead of Z to obtain

$$\begin{aligned}
&\mathbb{E}\Big[\mu\big(\sqrt{k^Z - k^Z_{s,X}}\big)\Big] \\
&\le 2\mathbb{E}\Big[\mu_X\big(\sqrt{k^Z_X - k^Z_{s,X}}\big)\Big] + 4\sqrt{\sum_{i>m} \sigma_i + \frac{\sqrt{k_{\max}}}{M}\left(\frac{80 m^2 \log(1+2M)}{9} + 69\right)}.
\end{aligned} \tag{5.36}$$

for any integer $m \ge 1$, where we have used $\|Pk(\cdot,x)\|_{\mathcal{H}_k} = \sqrt{k^Z(x,x) - k^Z_{s,X}(x,x)} = \sqrt{k^Z_X(x,x) - k^Z_{s,X}(x,x)}$ almost surely. From Proposition 5.13, we have

$$\mathbb{E}\Big[\mu_X\big(\sqrt{k^Z_X - k^Z_{s,X}}\big)\Big]^2 \le \mathbb{E}\Big[\mu_X\big(\sqrt{k^Z_X - k^Z_{s,X}}\big)^2\Big] \le \mathbb{E}\big[\mu_X(k^Z_X - k^Z_{s,X})\big] \le \sum_{i>s} \sigma_i,$$

and combining it with (5.36) leads to the desired conclusion. □

5.B.14 Proof of Theorem 5.16

Proof. We first prove the result for $Q_n = \mathrm{KQuad}(\mathrm{k_{s,Y}}, \mathrm{Y})$. Since $k(x,x) \ge k^Z(x,x) = k^Z_Y(x,x) \ge k^Y_{s,Z}(x,x)$ for $x \in Y$ from Proposition 5.12, we have

$$\mu_Y(\sqrt{k - k^Z_{s,Y}}) \le \mu_Y(\sqrt{k - k^Z}) + \mu_Y(\sqrt{k^Z_Y - k^Z_{s,Y}}).$$

From Proposition 5.13, by taking the expectation with regard to Y, we have

$$\mathbb{E}\Big[\mu_Y(\sqrt{k^Z_Y - k^Z_{s,Y}})\Big] \le \sqrt{\mathbb{E}\big[\mu_Y(k^Z_Y - k^Z_{s,Y})\big]} \le \sqrt{\sum_{i>s} \sigma_i},$$

and so we obtain

$$\mathbb{E}\Big[\mu_Y(\sqrt{k - k^Z_{s,Y}})\Big] \le \mu(\sqrt{k - k^Z}) + \sqrt{\sum_{i>s} \sigma_i}$$

By combining it with (5.13), it is now sufficient to show $\mathbb{E}[\mathrm{MMD}_k(\mu_Y, \mu)] \le \sqrt{c_{k,\mu}/N}$, but actually, it follows from the identity $\mathbb{E}[\mathrm{MMD}_k(\mu_Y, \mu)^2] = c_{k,\mu}/N$, which can be shown by a straightforward calculation (see, e.g., proof of Theorem 4.7).

In the case of $Q_n = \mathrm{KQuad}(k^Z_{s,\mu}, Y)$, we instead have the decomposition

$$\mu_Y(\sqrt{k - k^Z_{s,\mu}}) \le \mu_Y(\sqrt{k - k^Z}) + \mu_Y(\sqrt{k^Z_\mu - k^Z_{s,\mu}});$$

Theorem 5.11 yields the desired estimate for expectation. □

Chapter 6

Conclusion

We have studied the discretization of probability measures from two viewpoints. For random convex hulls, we have derived sharp bounds on the probability $p_{n,X}(\theta)$, which enables us to estimate the computational complexity of randomized construction of general cubature rules for a random vector X. For kernel quadrature, we have proposed an algorithm for constructing a kernel quadrature rule that enjoys (1) a practical algorithm with recombination and the Nyström approximation and (2) theoretical guarantees based on the spectral decay associated with the given kernel-measure pair.

To conclude it, we shall discuss the studies of Bayesian methods [1, 2, 3] obtained as applications of our research in Section 6.1 and other possible future research directions after this thesis in Section 6.2.

6.1 Bayesian numerical methods

As a direct application of our study, we work on Bayesian numerical methods. The setting is that we have an unknown function, which we want to estimate or optimize, modeled by a Gaussian process $f \sim \mathcal{GP}(m, k)$, where $m : \mathcal{X} \to \mathbb{R}$ is the mean function and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the covariance kernel (i.e., $\mathbb{E}[f(x)] = m(x)$, $\mathbb{E}[(f(x) - m(x))(f(y) - m(y))] = k(x, y)$).

There can be several regimes on the nature of the problem, but we assume that evaluating the function f is expensive but can be queried in parallel, such as simulation-based parameter estimation [113] or drug discovery [26]. In this setting, we observe functions values in a *batch* $X = (x_i)_{i=1}^n$; we observe n function values $\mathbf{y} = f(X) = (f(x_i))_{i=1}^n$ at the same time. Our task here is to find a good point set X to reduce the uncertainty as much as we can, i.e., to reduce the variance of the posterior GP $f \,|\, \mathbf{y} \sim \mathcal{GP}(m_{\mathbf{y}}, k_{\mathbf{y}})$ with

$$\begin{aligned} m_{\mathbf{y}}(x) &= m(x) + k(x, X)k(X, X)^{-1}\mathbf{y}, \\ k_{\mathbf{y}}(x, x') &= k(x, x') - k(x, X)k(X, X)^{-1}k(X, x'). \end{aligned}$$

In batch Bayesian quadrature/inference/optimization, we query the function values at an optimized batch X, update the GP as above, and repeat this iteration to reduce the uncertainty. In reality, this GP can be warped (e.g., we can only observe the square f^2, not f) and we also need to take into account the optimization and update of hyperparameters of the prior covariance k, but we here only consider the vanilla GP for simplicity.

In Adachi et al. [1], we address the problem of Bayesian quadrature [131], where we want to estimate an integral $Z = \int_{\mathcal{X}} f(x) \,\mathrm{d}\mu(x)$ for a Borel probability measure μ. From the linearity, the posterior $Z \,|\, \mathbf{y}$ is a Gaussian variable with

$$\mathbb{E}[Z \,|\, \mathbf{y}] = \int_X m_{\mathbf{y}}(x) \,\mathrm{d}\mu(x), \quad \mathbb{V}\mathrm{ar}[Z \,|\, \mathbf{y}] = \iint_{\mathcal{X}\times\mathcal{X}} k_{\mathbf{y}}(x, y) \,\mathrm{d}\mu(x) \,\mathrm{d}\mu(y).$$

This is closely related to kernel quadrature, that we discussed in the previous section. Let us consider the quadrature $Q_{\mathbf{w},X}$ given by weights $\mathbf{w} = (w_i)_{i=1}^n$ and points $X = (x_i)_{i=1}^n$. Then, the posterior variance after observing $\mathbf{y} = f(X)$ can be represented as the minimum worst-case error for the fixed point configuration X [78]:

$$\mathbb{V}\mathrm{ar}[Z \,|\, \mathbf{y}] = \inf_{\mathbf{w}\in\mathbb{R}^n} \mathrm{wce}(Q_{\mathbf{w},X}; \mathcal{H}_k, \mu)^2.$$

By using this equivalence, we can import our method in kernel quadrature to this iterative batch Bayesian quadrature. Although we need some engineering, e.g., when sampling from μ is not easy, our proposed method works as a scalable batch Bayesian quadrature as well as additional Bayesian inference.

In the more recent papers [2, 3], we approach the batch Bayesian optimization problem from the viewpoint of discretization of measures. When we have a function

f_{true} we want to maximize, it can be seen as optimization over $P(\mathcal{X})$, the set of all the Borel probability measures on $\mathcal{X}$:

$$\delta_{x^*} \in \operatorname*{arg\,max}_{\mu \in P(\mathcal{X})} \int_{\mathcal{X}} f_{\mathrm{true}}(x)\,\mathrm{d}\mu(x) \qquad \mathit{if} \qquad x^* \in \operatorname*{arg\,max}_{x\in\mathcal{X}} f_{\mathrm{true}}(x).$$

Our basic idea then is to iteratively update the probability measure μ representing the estimated location of the maximum x^* and efficiently discretize this "region of interest" by kernel/Bayesian quadrature to find an informative point configuration as the next batch sample. This can also be seen as an acceleration of the batch Thompson sampling [75], and shows a strong empirical performance in terms of sample efficiency as well as wall-clock computational time.

We believe that these are just a part of many possible applications of numerical quadrature to the field of Bayesian numerics. One problem with our approach is that, since Bayesian methods iteratively update the Gaussian process, we do not have a theoretical guarantee of the full quadrature/optimization process yet, whereas efficiency within one iteration can be explained by the spectral decay of the covariance kernel. In other words, we lose track of the change in spectral decay over iterations; this is an outstanding theoretical question.

6.2 Future directions

At last, we discuss a few directions for future research that can follow from this thesis.

Weighted sampling. When we construct a quadrature, what we often do is first sample a large number of candidate points from the target distribution μ and take its weighted subset for approximating μ, as described in Chapter 4.

However, in the setting of kernel quadrature with kernel k, it has been suggested that sampling from ν with $\mathrm{d}\nu(x) \propto \sqrt{k(x,x)}\,\mathrm{d}\mu(x)$ has better properties, such as smaller error of kernel mean embedding [1, 176]. Another example is that, when we are given an n-dimensional subspace $V_n \subset L^2(\mu)$ spanned by an orthonormal basis $\{e_1, \dots, e_n\}$, sampling from ν with $\mathrm{d}\nu(x) = n^{-1}(e_1(x)^2 + \cdots + e_n(x)^2)\,\mathrm{d}\mu(x)$ has favorable properties in terms of function approximation [31] and numerical integration [116] with respect to V_n.

In our context, it would be great to understand the behavior of the random convex hulls given by the above-mentioned (or more general) weighted sampling. In this way, we could construct a cubature with random convex hulls, even with a small Tukey depth $\alpha_X(\mathbb{E}[X])$.

Signature kernels. The initial motivation of hypercontractivity studies in Chapter 3 was on the possibility of random construction of the cubature on Wiener space [110] with general degree and dimension. Cubature on Winer space can be regarded as an exact kernel quadrature for a truncated signature kernel [91] with respect to the Wiener measure. We can naturally consider its untruncated version with an appropriate signature kernel and apply our method; the relation between signature kernel quadrature and cubature on Wiener space is also discussed in Cass et al. [28, Section 7.2]. So the interesting problem in this regard is the spectral decay of the integral operator (1.3) of the (lifted or unlifted) signature kernel with the Wiener measure.

Stochastic optimization. A typical setting of applications of stochastic optimization is that, we want to minimize a loss function $L(\theta) := \mathbb{E}[f_\theta(X)]$ with respect to $\theta \in \Theta$, where $X \sim \mu$ is random "data" taking values in $\mathcal{X}$ and $(f_\theta)_{\theta\in\Theta}$ is a parametrized family of functions $\mathcal{X} \to \mathbb{R}$. When μ is a massive discrete measure or an intractable continuous measure, by the so-called mini-batch stochastic gradient descent, we approximate the gradient $\nabla_\theta L(\theta)$ as

$$\nabla_\theta L(\theta) = \mathbb{E}[\partial_\theta f_\theta(X)] \approx \frac{1}{N}\sum_{i=1}^{N} \partial_\theta f_\theta(X_i),$$

where $X_1, \ldots, X_N$ are i.i.d. sample from μ.

This is basically a Monte Carlo estimate of the gradient, and it is reported that increasing the batch size N (i.e., having a more accurate gradient estimator) has a practical benefit in training high-quality GANs [23]. Ways to reduce the computational complexity (by carefully choosing a non-i.i.d. smaller batch) while keeping the accuracy of the gradient estimate are explored, e.g., coresets [155], DPPs [9], and Carathéodory subsampling [35]. If we choose an appropriate RKHS that fits $\partial_\theta f_\theta$, we could apply our kernel quadrature algorithms for updating the parameters.

Bibliography

[1] Masaki Adachi, Satoshi Hayakawa, Martin Jørgensen, Harald Oberhauser, and Michael A Osborne. Fast Bayesian inference with batch Bayesian quadrature via kernel recombination. In *Advances in Neural Information Processing Systems*, volume 35, pages 16533–16547, 2022.

[2] Masaki Adachi, Satoshi Hayakawa, Saad Hamid, Martin Jørgensen, Harald Oberhauser, and Micheal A Osborne. SOBER: Highly parallel Bayesian optimization and bayesian quadrature over discrete and mixed spaces. *arXiv preprint arXiv:2301.11832*, 2023.

[3] Masaki Adachi, Satoshi Hayakawa, Xingchen Wan, Martin Jørgensen, Harald Oberhauser, and Michael A Osborne. Domain-agnostic batch Bayesian optimization with diverse constraints via Bayesian quadrature. *arXiv preprint arXiv:2306.05843*, 2023.

[4] Mark Yuying An. Log-concave probability distributions: Theory and statistical testing. *Duke University Dept of Economics Working Paper*, 1997.

[5] Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E. Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, Lester Mackey, Chris J. Oates, Gesine Reinert, and Yvik Swan. Stein's Method Meets Computational Statistics: A Review of Some Recent Developments. *Statistical Science*, 38(1):120–139, 2023.

[6] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1): 714–751, 2017.

[7] Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *International Conference on Machine Learning*, pages 1355–1362, 2012.

[8] Keith Ball. The reverse isoperimetric problem for Gaussian measure. *Discrete & Computational Geometry*, 10(4):411–420, 1993.

[9] Rémi Bardenet, Subhroshekhar Ghosh, and Meixia Lin. Determinantal point processes based on orthogonal polynomials for sampling minibatches in SGD. In *Advances in Neural Information Processing Systems*, volume 34, pages 16226–16237, 2021.

[10] Alexander Barvinok. Thrifty approximations of convex bodies by polytopes. *International Mathematics Research Notices*, 2014(16):4341–4356, 2014.

[11] Fabrice Baudoin and Laure Coutin. Operators associated with a stochastic differential equation driven by fractional Brownian motions. *Stochastic Processes and their Applications*, 117(5):550–574, 2007.

[12] Christian Bayer and Josef Teichmann. The proof of Tchakaloff's theorem. *Proceedings of the American Mathematical Society*, 134(10):3035–3040, 2006.

[13] William Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, 102(1):159–182, 1975.

[14] William Beckner. Sobolev inequalities, the Poisson semigroup, and analysis on the sphere s^n. *Proceedings of the National Academy of Sciences*, 89(11): 4816–4819, 1992.

[15] Ayoub Belhadji. An analysis of Ermakov–Zolotukhin quadrature using kernels. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[16] Ayoub Belhadji, R Bardenet, and Pierre Chainais. Kernel quadrature with DPPs. In *Advances in Neural Information Processing Systems*, volume 32, pages 12907–12917, 2019.

[17] Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. Kernel interpolation with continuous volume sampling. In *International Conference on Machine Learning*, pages 725–735. PMLR, 2020.

[18] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer, 2004.

[19] Andrew C Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941.

[20] Aline Bonami. Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Annales de l'institut Fourier*, 20(2):335–402, 1970.

[21] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[22] François-Xavier Briol, Chris Oates, Mark Girolami, and Michael A Osborne. Frank–Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. *Advances in Neural Information Processing Systems*, 28:1162–1170, 2015.

[23] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

[24] Andrew Caplin and Barry Nalebuff. Aggregation and social choice: A mean voter theorem. *Econometrica: Journal of the Econometric Society*, pages 1–23, 1991.

[25] Constantin Carathéodory. Über den variabilitätsbereich der fourier'schen konstanten von positiven harmonischen funktionen. *Rendiconti Del Circolo Matematico di Palermo (1884-1940)*, 32(1):193–217, 1911.

[26] Arnaud Carpentier, Ila Nimgaonkar, Virginia Chu, Yuchen Xia, Zongyi Hu, and T Jake Liang. Hepatic differentiation of human pluripotent stem cells in miniaturized format suitable for high-throughput screen. *Stem Cell Research*, 16(3):640–650, 2016.

[27] Ignacio Cascos. Depth functions based on a number of observations of a random vector. *Working paper 07–07, Statistics and Econometrics Series, Universidad Carlos III de Madrid*, 2007. URL `https://hdl.handle.net/10016/700`.

[28] Thomas Cass, Terry Lyons, and Xingcheng Xu. General signature kernels. *arXiv preprint arXiv:2107.00447*, 2021.

[29] Antoine Chatalic, Nicolas Schreuder, Lorenzo Rosasco, and Alessandro Rudi. Nyström kernel mean embeddings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 3006–3024, 2022.

[30] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Conference on Uncertainty in Artificial Intelligence*, pages 109–116, 2010.

[31] Albert Cohen and Giovanni Migliorati. Optimal weighted least-squares methods. *The SMAI journal of computational mathematics*, 3:181–203, 2017.

[32] Cyrille W Combettes and Sebastian Pokutta. Revisiting the approximate Carathéodory problem via the Frank–Wolfe algorithm. *Mathematical Programming*, 197(1):191–214, 2023.

[33] John B Conway. *A course in functional analysis*. Springer, 2007.

[34] Francesco Cosentino, Harald Oberhauser, and Alessandro Abate. A randomized algorithm to reduce the support of discrete measures. In *Advances in Neural Information Processing Systems*, volume 33, pages 15100–15110, 2020.

[35] Francesco Cosentino, Harald Oberhauser, and Alessandro Abate. Carathéodory sampling for stochastic gradient descent. *arXiv preprint arXiv:2006.01819*, 2020.

[36] Juan Antonio Cuesta-Albertos and Alicia Nieto-Reyes. The random Tukey depth. *Computational Statistics & Data Analysis*, 52(11):4979–4988, 2008.

[37] N. Dafnis, A. Giannopoulos, and A. Tsolomitis. Asymptotic shape of a random polytope in a convex body. *Journal of Functional Analysis*, 257(9): 2820–2839, 2009.

[38] S. De Marchi. On optimal center locations for radial basis function interpolation: computational aspects. *Rendiconti del Seminario Matematico*, 61(3): 343–358, 2003.

[39] Stefano De Marchi, Robert Schaback, and Holger Wendland. Near-optimal data-independent point locations for radial basis function interpolation. *Advances in Computational Mathematics*, 23(3):317–330, 2005.

[40] Philipp J di Dio and Konrad Schmüdgen. The multidimensional truncated moment problem: the moment cone. *Journal of Mathematical Analysis and Applications*, 511(1):126066, 2022.

[41] Josef Dick, Frances Y. Kuo, and Ian H. Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

[42] David L Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803–1827, 1992.

[43] Petros Drineas, Michael W Mahoney, and Nello Cristianini. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6(12):2153–2175, 2005.

[44] Raaz Dwivedi and Lester Mackey. Kernel thinning. In *Conference on Learning Theory*, pages 1753–1753. PMLR, 2021.

[45] Raaz Dwivedi and Lester Mackey. Generalized kernel thinning. In *International Conference on Learning Representations*, 2022.

[46] Ethan N Epperly and Elvira Moreno. Kernel quadrature with randomly pivoted cholesky. *arXiv preprint arXiv:2306.03955*, 2023.

[47] Alexandros Eskenazis, Piotr Nayar, and Tomasz Tkocz. Sharp comparison of moments and the log-concave moment problem. *Advances in Mathematics*, 334:389–416, 2018.

[48] Carl-Gustav Esseen. On the Liapunoff limit of error in the theory of probability. *Arkiv for Matematik, Astronomi och Fysik, A: 1–19*, 1942.

[49] Gregory E. Fasshauer and Michael J. McCourt. Stable evaluation of Gaussian radial basis function interpolants. *SIAM Journal on Scientific Computing*, 34(2):A737–A762, 2012. URL `https://doi.org/10.1137/110824784`.

[50] Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

[51] Bertrand Gauthier. Nyström approximation and reproducing kernels: embeddings, projections and squared-kernel discrepancy. *preprint*, 2021. URL `https://hal.archives-ouvertes.fr/hal-03207443`.

[52] Apostolos Giannopoulos and Marianna Hartzoulaki. Random spaces generated by vertices of the cube. *Discrete and Computational Geometry*, 28(2): 255–273, 2002.

[53] Evarist Giné and Vladimir Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

[54] Alex Gittens and Michael W Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *The Journal of Machine Learning Research*, 17(1):3977–4041, 2016.

[55] Jan Glaubitz. Stable high-order cubature formulas for experimental data. *Journal of Computational Physics*, 447:110693, 2021.

[56] E D Gluskin. Extremal properties of orthogonal parallelepipeds and their applications to the geometry of Banach spaces. *Mathematics of the USSR-Sbornik*, 64(1):85–96, 1989.

[57] Jun-ya Gotoh, Akiko Takeda, and Katsuya Tono. DC formulations and algorithms for sparse optimization problems. *Mathematical Programming*, 169(1):141–176, 2018.

[58] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[59] Branko Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. *Pacific Journal of Mathematics*, 10(4):1257–1261, 1960.

[60] Olivier Guédon, Felix Krahmer, Christian Kümmerle, Shahar Mendelson, and Holger Rauhut. On the geometry of polytopes generated by heavy-tailed random vectors. *Communications in Contemporary Mathematics*, 24 (03):2150056, 2022.

[61] O. Guédon, A. E. Litvak, and K. Tatarko. Random polytopes obtained by matrices with heavy-tailed entries. *Communications in Contemporary Mathematics*, 22(04):1950027, 2020.

[62] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2006.

[63] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[64] John H Halton. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, 2(1):84–90, 1960.

[65] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 1952.

[66] Satoshi Hayakawa. Monte Carlo cubature construction. *Japan Journal of Industrial and Applied Mathematics*, 38:561–577, 2021.

[67] Satoshi Hayakawa and Taiji Suzuki. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, 2020.

[68] Satoshi Hayakawa and Ken'ichiro Tanaka. Monte Carlo construction of cubature on Wiener space. *Japan Journal of Industrial and Applied Mathematics*, 39(2):543–571, 2022.

[69] Satoshi Hayakawa and Ken'ichiro Tanaka. Convergence analysis of approximation formulas for analytic functions via duality for potential energy minimization. *Japan Journal of Industrial and Applied Mathematics*, 2023.

[70] Satoshi Hayakawa, Harald Oberhauser, and Terry Lyons. Positively weighted kernel quadrature via subsampling. In *Advances in Neural Information Processing Systems*, volume 35, pages 6886–6900, 2022.

[71] Satoshi Hayakawa, Terry Lyons, and Harald Oberhauser. Estimating the probability that a given vector is in the convex hull of a random sample. *Probability Theory and Related Fields*, 185(3-4):705–746, 2023.

[72] Satoshi Hayakawa, Harald Oberhauser, and Terry Lyons. Hypercontractivity meets random convex hulls: analysis of randomized multivariate cubatures. *Proceedings of the Royal Society A*, 479(2273):20220725, 2023.

[73] Satoshi Hayakawa, Harald Oberhauser, and Terry Lyons. Sampling-based Nyström approximation and kernel quadrature. In *Proceedings of the 40th International Conference on Machine Learning*, pages 12678–12699, 2023.

[74] Pierre Henry-Labordere, Xiaolu Tan, and Nizar Touzi. Unbiased simulation of stochastic differential equations. *The Annals of Applied Probability*, 27 (6):3305–3341, 2017.

[75] José Miguel Hernández-Lobato, James Requeima, Edward O Pyzer-Knapp, and Alán Aspuru-Guzik. Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space. In *International conference on machine learning*, pages 1470–1479. PMLR, 2017.

[76] J Ben Hough, Manjunath Krishnapur, Yuval Peres, Bálint Virág, et al. Determinantal processes and independence. *Probability surveys*, 3:206–229, 2006.

[77] Daniel Hug. Random polytopes. In *Stochastic geometry, spatial statistics and random fields*, pages 205–238. Springer, 2013.

[78] Ferenc Huszár and David Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *Conference on Uncertainty in Artificial Intelligence*, pages 377–386, 2012.

[79] Svante Janson. *Gaussian Hilbert spaces*. Cambridge University Press, 1997.

[80] Rong Jin, Tianbao Yang, Mehrdad Mahdavi, Yu-Feng Li, and Zhi-Hua Zhou. Improved bounds for the Nyström method with application to kernel classification. *IEEE Transactions on Information Theory*, 59(10):6939–6949, 2013.

[81] Zakhar Kabluchko and Dmitry Zaporozhets. Absorption probabilities for Gaussian polytopes and regular spherical simplices. *Advances in Applied Probability*, 52(2):588—616, 2020.

[82] Motonobu Kanagawa, Bharath K Sriperumbudur, and Kenji Fukumizu. Convergence guarantees for kernel-based quadrature rules in misspecified settings. *Advances in Neural Information Processing Systems*, 29:3296–3304, 2016.

[83] Motonobu Kanagawa, Bharath K Sriperumbudur, and Kenji Fukumizu. Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *Foundations of Computational Mathematics*, 20(1):155–194, 2020.

[84] Toni Karvonen and Simo Särkkä. Gaussian kernel quadrature at scaled Gauss–Hermite nodes. *BIT Numerical Mathematics*, 59(4):877–902, 2019.

[85] Toni Karvonen, Chris J Oates, and Simo Särkkä. A Bayes–Sard cubature method. In *Advances in Neural Information Processing Systems*, volume 31, pages 5886–5897, 2018.

[86] Toni Karvonen, Chris Oates, and Mark Girolami. Integration in reproducing kernel hilbert spaces of Gaussian kernels. *Mathematics of Computation*, 90 (331):2209–2233, 2021.

[87] Toni Karvonen, Simo Särkkä, and Ken'ichiro Tanaka. Kernel-based interpolation at approximate Fekete points. *Numerical Algorithms*, 87(1):445–468, 2021.

[88] Manohar Kaul, Bin Yang, and Christian S Jensen. Building accurate 3d spatial networks to enable next generation intelligent transportation systems. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 1, pages 137–146. IEEE, 2013.

[89] Heysem Kaya, Pmar Tüfekci, and Fikret S Gürgen. Local and global learning methods for predicting power of a combined gas & steam turbine. In *Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering*, pages 13–18, 2012.

[90] Jeong Han Kim and Van H Vu. Concentration of multivariate polynomials and its applications. *Combinatorica*, 20(3):417–434, 2000.

[91] Franz J. Kiraly and Harald Oberhauser. Kernels for sequentially ordered data. *The Journal of Machine Learning Research*, 20(31):1–45, 2019.

[92] Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[93] Vladimir Koltchinskii and Evarist Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000.

[94] Hermann König. On the best constants in the Khintchine inequality for Steinhaus variables. *Israel Journal of Mathematics*, 203(1):23–57, 2014.

[95] Victor Korolev and Irina Shevtsova. An improvement of the Berry-Esseen inequality with applications to Poisson and mixed Poisson random sums. *Scandinavian Actuarial Journal*, 2012(2):81–105, 2012.

[96] David Krieg and Mathias Sonnleitner. Random points are optimal for the approximation of sobolev functions. *IMA Journal of Numerical Analysis*, 2023.

[97] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the Nyström method. *The Journal of Machine Learning Research*, 13(1): 981–1006, 2012.

[98] Anastasios Kyrillidis, Stephen Becker, Volkan Cevher, and Christoph Koch. Sparse projections onto the simplex. In *International Conference on Machine Learning*, pages 235–243. PMLR, 2013.

[99] FM Larkin. Optimal approximation in Hilbert spaces with reproducing kernel functions. *Mathematics of Computation*, 24(112):911–921, 1970.

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

[101] Mu Li, Wei Bi, James T Kwok, and Bao-Liang Lu. Large-scale Nyström kernel matrix approximation using randomized SVD. *IEEE Transactions on Neural Networks and Learning Systems*, 1(26):152–164, 2015.

[102] Ping Li, Syama Sundar Rangapuram, and Martin Slawski. Methods for sparse and low-rank recovery under simplex constraints. *Statistica Sinica*, 30(2):557–577, 2020.

[103] C. Litterer and T. Lyons. High order recombination and an application to cubature on Wiener space. *The Annals of Applied Probability*, 22(4):1301–1327, 2012.

[104] A.E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Advances in Mathematics*, 195(2):491–523, 2005.

[105] Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2021.

[106] Qiang Liu and Jason Lee. Black-box importance sampling. In *Artificial Intelligence and Statistics*, pages 952–961. PMLR, 2017.

[107] Regina Y. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

[108] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[109] Junwei Lu, Guang Cheng, and Han Liu. Nonparametric heterogeneity testing for massive data. *arXiv preprint arXiv:1601.06212v1*, 2016.

[110] Terry Lyons and Nicolas Victoir. Cubature on Wiener space. *Proceedings of the Royal Society of London Series A*, 460:169–198, 2004.

[111] Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8305–8316, 2019.

[112] Satya N Majumdar, Alain Comtet, and Julien Randon-Furling. Random convex hulls and extreme value statistics. *Journal of Statistical Physics*, 138 (6):955–1009, 2010.

[113] S. G. Marquis, V. Sulzer, R. Timms, C. P. Please, and S. J. Chapman. An asymptotic derivation of a single particle model with electrolyte. *Journal of the Electrochemical Society*, 166(15):A3693–A3706, 2019.

[114] John C Mason and David C Handscomb. *Chebyshev polynomials*. CRC press, 2002.

[115] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 2021.

[116] Giovanni Migliorati and Fabio Nobile. Stable high-order randomized cubature formulae in arbitrary dimension. *Journal of Approximation Theory*, page 105706, 2022.

[117] Ha Quang Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2010.

[118] Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pages 154–168. Springer, 2006.

[119] Vahab Mirrokni, Renato Paes Leme, Adrian Vladu, and Sam Chiu-wai Wong. Tight bounds for approximate Carathéodory and beyond. In *International Conference on Machine Learning*, pages 2440–2448. PMLR, 2017.

[120] Shinji Mizuno. Polynomiality of infeasible-interior-point algorithms for linear programming. *Mathematical Programming*, 67(1-3):109–119, 1994.

[121] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018.

[122] Karl Mosler. Depth statistics. In *Robustness and complex data structures*, pages 17–34. Springer, 2013.

[123] 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(1-2):1–141, 2017.

[124] Stanislav Nagy, Carsten Schütt, and Elisabeth M. Werner. Halfspace depth and floating body. *Statistics Surveys*, 13(none):52 – 118, 2019.

[125] Edward Nelson. The free Markoff field. *Journal of Functional Analysis*, 12 (2):211–227, 1973.

[126] Syoiti Ninomiya and Yuji Shinozaki. On implementation of high-order recombination and its application to weak approximations of stochastic differential equations. In *Proceedings of the NFA 29th Annual Conference*, 2021.

[127] Ivan Nourdin, Giovanni Peccati, and Gesine Reinert. Invariance principles for homogeneous sums: universality of Gaussian Wiener chaos. *The Annals of Probability*, 38(5):1947–1985, 2010.

[128] Erich Novak. *Deterministic and stochastic error bounds in numerical analysis*. Springer, 1988.

[129] CJ Oates, M Girolami, and N Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 79:695–718, 2017.

[130] Dino Oglic and Thomas Gärtner. Nyström method with kernel k-means++ samples as landmarks. In *International Conference on Machine Learning*, pages 2652–2660. PMLR, 2017.

[131] Anthony O'Hagan. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, 1991.

[132] Art B Owen. A randomized Halton algorithm in R. *arXiv preprint arXiv:1706.02808*, 2017.

[133] V. Pan. On the complexity of a pivot step of the revised simplex algorithm. *Computers & Mathematics with Applications*, 11(11):1127 – 1140, 1985.

[134] Riccardo Passeggeri. Some results on the signature and cubature of the fractional Brownian motion for $H > \frac{1}{2}$. *arXiv preprint arXiv:1609.07352*, 2016.

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

[136] Roger Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

[137] Mert Pilanci, Laurent El Ghaoui, and Venkat Chandrasekaran. Recovery of sparse probability measures via convex programming. In *Advances in Neural Information Processing Systems*, volume 25, pages 2420–2428, 2012.

[138] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

[139] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184, 2007.

[140] Martin Raič. A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli*, 25(4A):2824–2853, 2019.

[141] Hans Richter. Parameterfreie abschätzung und realisierung von erwartungswerten. *Blätter der DGVFM*, 3(2):147–162, 1957.

[142] Werner Wolfgang Rogosinski. Moments of non-negative mass. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 245(1240):1–27, 1958.

[143] Paul C Rosenbloom. Quelques classes de problèmes extrémaux. ii. *Bulletin de la societe mathematique de France*, 80:183–215, 1952.

[144] Peter J Rousseeuw and Ida Ruts. The depth function of a population distribution. *Metrika*, 49(3):213–244, 1999.

[145] Gabriele Santin and Bernard Haasdonk. Convergence rate of the data-independent P-greedy algorithm in kernel-based approximation. *Dolomites Research Notes on Approximation*, 10, 2017.

[146] Gabriele Santin and Robert Schaback. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42(4):973–993, 2016.

[147] Arthur Sard. Best approximate integration formulas; best approximation formulas. *American Journal of Mathematics*, 71(1):80–91, 1949.

[148] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

[149] Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 437–446. SIAM, 2012.

[150] Carsten Schütt and Elisabeth Werner. The convex floating body. *Mathematica Scandinavica*, 66(2):275–290, 1990.

[151] Ron Shamir. The efficiency of the simplex method: A survey. *Management Science*, 33(3):301–334, 1987.

[152] Abhishek Shetty, Raaz Dwivedi, and Lester Mackey. Distribution compression in near-linear time. In *International Conference on Learning Representations*, 2022.

[153] Nobuo Shinozaki, Masaaki Sibuya, and Kunio Tanabe. Numerical algorithms for the Moore–Penrose inverse of a matrix: iterative methods. *Annals of the Institute of Statistical Mathematics*, 24(1):621–629, 1972.

[154] Barry Simon and Raphael Høegh-Krohn. Hypercontractive semigroups and two dimensional self-coupled Bose fields. *Journal of Functional Analysis*, 9 (2):121–180, 1972.

[155] Samarth Sinha, Han Zhang, Anirudh Goyal, Yoshua Bengio, Hugo Larochelle, and Augustus Odena. Small-GAN: Speeding up GAN training using core-sets. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9005–9015. PMLR, 2020.

[156] Leah F South, Toni Karvonen, Chris Nemeth, Mark Girolami, and Chris J Oates. Semi-exact control functionals from Sard's method. *Biometrika*, 109 (2):351–367, 2022.

[157] Nathan Srebro and Karthik Sridharan. Note on refined Dudley integral covering number bound. *Unpublished results*, 2010. URL `https://www.cs.cornell.edu/~sridharan/dudley.pdf`.

[158] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

[159] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11: 1517–1561, 2010.

[160] Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

[161] Arthur H Stroud. *Approximate calculation of multiple integrals.* Prentice-Hall, 1971.

[162] V. Tchakaloff. Formules de cubature mécanique à coefficients non négatifs. *Bulletin des Sciences Mathématiques*, 81:123–134, 1957.

[163] Maria Tchernychova. *Carathéodory cubature measures.* PhD thesis, University of Oxford, 2016.

[164] Onur Teymur, Jackson Gorham, Marina Riabiz, and Chris Oates. Optimal quantisation of probability measures using maximum mean discrepancy. In *International Conference on Artificial Intelligence and Statistics*, pages 1027–1035. PMLR, 2021.

[165] Anthony Tompkins and Fabio Ramos. Fourier feature approximations for periodic kernels in time-series modelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[166] Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Fixed-rank approximation of a positive-semidefinite matrix from streaming data. *Advances in Neural Information Processing Systems*, 30, 2017.

[167] Kazuma Tsuji, Ken'ichiro Tanaka, and Sebastian Pokutta. Pairwise conditional gradients without swap steps and sparser kernel herding. In *International Conference on Machine Learning*, pages 21864–21883. PMLR, 2022.

[168] Pınar Tüfekci. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60:126–140, 2014.

[169] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

[170] Paxton Turner, Jingbo Liu, and Philippe Rigollet. A statistical perspective on coreset density estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 2512–2520. PMLR, 2021.

[171] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[172] Uli Wagner and Emo Welzl. A continuous analogue of the upper bound theorem. *Discrete & Computational Geometry*, 26(2):205–219, 2001.

[173] Grace Wahba. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, 1990.

[174] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[175] Abraham Wald. Limits of a distribution function determined by absolute moments and inequalities satisfied by absolute moments. *Transactions of the American Mathematical Society*, 46(2):280–306, 1939.

[176] Congye Wang, Wilson Chen, Heishiro Kanagawa, and Chris Oates. Stein π-importance sampling. *arXiv preprint arXiv:2305.10068*, 2023.

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

[178] J. G. Wendel. A problem in geometric probability. *Mathematica Scandinavica*, 11(1):109–111, 1963.

[179] Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, pages 661–667, 2000.

[180] Hayata Yamasaki, Sathyawageeswar Subramanian, Satoshi Hayakawa, and Sho Sonoda. Quantum ridgelet transform: Winning lottery ticket of neural networks with quantum computation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 39008–39034, 2023.

[181] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random Fourier features: A theoretical and empirical comparison. *Advances in Neural Information Processing Systems*, 25, 2012.

[182] Alex Zhai. A high-dimensional CLT in $\mathcal{W}_2$ distance with near optimal convergence rate. *Probability Theory and Related Fields*, 170(3-4):821–845, 2018.

[183] Alexander Alipkanovich Zhensykbaev. Monosplines of minimal norm and the best quadrature formulae. *Russian Mathematical Surveys*, 36(4):121–180, 1981.

[184] Yijun Zuo. A new approach for the computation of halfspace depth in high dimensions. *Communications in Statistics-Simulation and Computation*, 48 (3):900–921, 2019.

[185] Barbara Zwicknagl. Power series kernels. *Constructive Approximation*, 29: 61–84, 2009.