
seit 1558

# High-Dimensional Function Approximation: Breaking the Curse with Monte Carlo Methods

**D I S S E R T A T I O N**

*zur Erlangung des akademischen Grades*
*doctor rerum naturalium (Dr. rer. nat.)*

vorgelegt dem Rat der
Fakultät für Mathematik und Informatik
der Friedrich-Schiller-Universität Jena

von M.Sc. Robert Kunsch
geboren am 19. Februar 1990 in Gotha

**Gutachter**

1. Prof. Dr. E. Novak
2. Prof. Dr. H. Woźniakowski, Warschau
3. Prof. Dr. A. Hinrichs, Linz

**Tag der öffentlichen Verteidigung**: 28. Juli 2017

# Acknowledgements

First of all, I wish to express my deepest gratitude to my supervisor Prof. Dr. Erich Novak. His sincere advice, not only on scientific matters, carried me forward in so many ways during my time as his student. He helped me to realize when attempts to solve a problem might be unsuccesful. Even more often his encouragement led me to pursue new ideas. Not to forget, the numerous hints on the written text allowed the dissertation to evolve.

I am also grateful to my many colleagues for all their support. Especially David Krieg, who meticulously read large portions of the manuscript. Hopefully, his remarks led to a more comprehensive text. I would also like to thank Prof. Dr. Stefan Heinrich, Prof. Dr. Aicke Hinrichs, Prof. Dr. Winfried Sickel, Prof. Dr. Henryk Woźniakowski, Dr. Daniel Rudolf, Dr. Mario Ullrich, Dr. Tino Ullrich, Dr. Jan Vybíral, and my fellow students Glenn Byrenheid and Dr. Van Kien Nguyen, for interesting and encouraging scientific discussions during conferences and seminars. If I were to include all the friendly associates at mathematical meetings, this list would extend considerably.

Finally, I wish to thank my friends Nico Kennedy and Rafael Silveira for reading the english introduction for linguistic aspects, never deterred from the formulas, even as non-mathematicians. I almost forgot to mention Kendra Horner who helped polish this very text.

# Zusammenfassung

Für viele Probleme, die in wissenschaftlich-technischen Anwendungen auftreten, ist es praktisch unmöglich, exakte Lösungen zu finden. Stattdessen sucht man Näherungslösungen mittels Verfahren, die in endlich vielen Schritten umsetzbar sind. Insbesondere muss man mit unvollständiger Information über die jeweilige Probleminstanz arbeiten; abgesehen von Strukturannahmen (dem sogenannten *a priori*-Wissen) können wir nur endlich viel Information sammeln, typischerweise in Form von $n$ reellen Zahlen aus Messungen oder vom Nutzer bereitzustellenden Unterprogrammen. Es besteht wachsendes Interesse an der Lösung hochdimensionaler Probleme, das sind Probleme mit Funktionen, die auf $d$-dimensionalen Gebieten definiert sind. Wir untersuchen die *Informationskomplexität* $n(\varepsilon, d)$, die minimal benötigte Anzahl von Informationen, um ein Problem bis auf einen Fehler $\varepsilon > 0$ zu lösen. *Tractability* ist die Frage nach dem Verhalten dieser Funktion $n(\varepsilon, d)$, also nach der *Durchführbarkeit* einer Aufgabe. In vielen Fällen wächst die Komplexität exponentiell in $d$ für festgehaltenes $\varepsilon$, wir sprechen dann vom *Fluch der Dimension*. Es gibt im Grunde zwei Wege, dem Fluch der Dimension zu begegnen – so er denn auftritt. Die eine Variante ist, mehr *a priori*-Wissen einzubeziehen und so die Menge der denkbaren Eingabegrößen einzugrenzen. Die andere Möglichkeit besteht in der Erweiterung der Klasse zulässiger Algorithmen. In dieser Dissertation liegt der Fokus auf dem zweiten Ansatz, und zwar untersuchen wir das Potential von Randomisierung für die Approximation von Funktionen.

Ein $d$-dimensionales Approximationsproblem ist eine Einbettungsabbildung

$$\mathrm{APP} : F^d \hookrightarrow G^d, \quad f \mapsto f\,,$$

mit einer *Inputmenge* $F^d$, die $d$-variate reellwertige Funktionen enthält, und einem normierten Raum $G^d$. Funktionen aus $F^d$ sind in dieser Arbeit für gewöhnlich auf dem $d$-dimensionalen Einheitswürfel $[0,1]^d$ definiert, der Zielraum $G^d$ ist dann beispielsweise $L_1([0,1]^d)$ oder $L_\infty([0,1]^d)$. Deterministische Algorithmen sind Abbildungen $A_n = \phi \circ N : F^d \to G^d$, wobei die *Informationsabbildung*

$$N : F^d \to \mathbb{R}^n, \quad f \mapsto (L_1(f), \ldots, L_n(f))\,,$$

mittels $n$ linearer Funktionale $L_i$ Information über die Probleminstanz $f$ sammelt, und $\phi : \mathbb{R}^n \to G^d$. Der Fehler des Verfahrens ist durch den ungünstigsten Fall bestimmt,

$$e(A_n, F^d) := \sup_{f \in F^d} \|f - A_n(f)\|_{G^d}\,.$$

Randomisierte Verfahren sind Familien $(A_n^\omega)_{\omega\in\Omega}$ von Abbildungen $A_n^\omega : F^d \to G^d$ obiger Struktur, indiziert durch ein Zufallselement $\omega \in \Omega$ aus einem Wahrscheinlichkeitsraum $(\Omega, \Sigma, \mathbb{P})$. Der Fehler eines solchen *Monte-Carlo*-Algorithmus ist der

erwartete Fehler für den schlechtesten Input,

$$e((A_n^\omega)_\omega, F^d) := \sup_{f \in F^d} \mathbb{E}\, \|f - A_n^\omega(f)\|_{G^d}\,.$$

In beiden Fällen wird eine signifikante Verkleinerung des *Anfangsfehlers* angestrebt,

$$e(0, F^d) := \inf_{g \in G^d} \sup_{f \in F^d} \|f - g\|_{G^d}\,,$$

welcher bereits ohne Information erreichbar ist. Insbesondere interessieren wir uns für den Vergleich der Komplexität im deterministischen und randomisierten Fall,

$$\begin{aligned} n^{\mathrm{det}}(\varepsilon, d) &:= \inf\{n \in \mathbb{N}_0 \mid \exists A_n : e(A_n, F^d) \le \varepsilon\}\,, \\ n^{\mathrm{ran}}(\varepsilon, d) &:= \inf\{n \in \mathbb{N}_0 \mid \exists (A_n^\omega)_\omega : e((A_n^\omega)_\omega, F^d) \le \varepsilon\}\,, \end{aligned}$$

wobei $0 < \varepsilon < e(0, F^d)$. Sämtliche in dieser Arbeit angegebene Algorithmen sind *nichtadaptiv* mit der einfachen Struktur von $N$ siehe oben. Untere Fehlerschranken werden für wesentlich allgemeinere Verfahren gezeigt, welche die Information auch *adaptiv* sammeln oder eine *veränderliche Kardinalität* $n(\omega, f)$ aufweisen. Zu diesen Begriffen und einer ausführlichen Einführung in das Themengebiet der *Informationskomplexität*, siehe Kapitel 1.

Neue Resultate sind in den Kapiteln 2–4 enthalten, welche mehr oder weniger für sich stehende Themen behandeln. Kapitel 2 befasst sich mit unteren Schranken für randomisierte Verfahren, mittels derer sich für verschiedene Beispiele zeigen lässt, dass Monte-Carlo-Methoden nicht viel besser als deterministische Algorithmen sein können. Im Gegensatz dazu ist Kapitel 3 der Suche nach Problemen gewidmet, wo deterministische Algorithmen unter dem Fluch der Dimension leiden, Randomisierung diesen jedoch auf recht eindrucksvolle Weise zu brechen vermag. Kapitel 4 bespricht ein konkretes Problem für welches Zufallsalgorithmen zwar den Fluch aufheben, das Problem aber trotzdem noch sehr schwer ist.

### Zu Kapitel 2: Untere Schranken für lineare Probleme mittels Bernstein-Zahlen

Das Hauptergebnis dieses Kapitels stellen untere Schranken für den Fehler von Monte-Carlo-Algorithmen für allgemeine lineare Probleme

$$S : F \to G$$

dar, d.h. $S$ ist ein linearer Operator zwischen normierten Räumen $\widetilde{F}$ und $G$, zudem ist die Inputmenge $F$ die Einheitskugel in $\widetilde{F}$. Es wird gezeigt, dass für jede (adaptive) Monte-Carlo-Methode $(A_n^\omega)_\omega$, welche $n$ beliebige stetige lineare Funktionale $L_i$ zur Informationsgewinnung einsetzt, die Abschätzung

$$e((A_n^\omega)_\omega, F) \ge \tfrac{1}{30}\, b_{2n}(S)$$

gilt, wobei $b_m(S)$ die $m$-te Bernstein-Zahl des Operators $S$ ist, siehe Theorem 2.1. Der Beweis basiert auf einem Ergebnis von Heinrich [22], welches Normerwartungswerte

von Gauß-Maßen in Beziehung zum Monte-Carlo-Fehler setzt. Die Neuerung besteht in der Anwendung des Theorems von Lewis für die Wahl eines optimalen Gauß-Maßes. Dieses Ergebnis wurde in [39] angekündigt und ein kurzer Beweis ohne explizite Konstanten aufgeführt.

In Abschnitt 2.4.2 wird dieses allgemeine Werkzeug für die $L_\infty$-Approximation bestimmter Klassen von $C^\infty$-Funktionen angewandt. Wir betrachten das Problem

$$\mathrm{APP} : F_p^d \hookrightarrow L_\infty([0,1]^d)$$

mit der Inputmenge

$$F_p^d := \{f \in C^\infty([0,1]^d) \mid \|\nabla_{\mathbf{v}_k} \cdots \nabla_{\mathbf{v}_1} f\|_\infty \le |\mathbf{v}_1|_p \cdots |\mathbf{v}_k|_p \\ \text{für } k \in \mathbb{N}_0,\, \mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^d\},$$

wobei $\nabla_{\mathbf{v}} f$ die Richtungsableitung entlang eines Vektors $\mathbf{v} \in \mathbb{R}^d$ bezeichnet und $|\mathbf{v}|_p$ die $p$-Norm von $\mathbf{v} \in \ell_p^d$, $1 \le p \le \infty$. Über die dazugehörigen Bernstein-Zahlen erhalten wir die untere Schranke

$$n^{\mathrm{ran}}(\varepsilon, d, p) > 2^{\lfloor \frac{d^{1/p}}{3} \rfloor - 1} \quad \text{für } 0 < \varepsilon \le \tfrac{1}{30},$$

siehe Corollary 2.20. Für $p = 1$ ergibt sich daraus der Fluch der Dimension, auch bei Randomisierung. Die Beweistechnik zur Bestimmung der Bernstein-Zahlen ist von Novak und Woźniakowski [56] bekannt, welche den Fluch der Dimension für den Fall $p = 1$ im deterministischen Szenario gezeigt haben.

Eine einfache Taylor-Approximation liefert obere Schranken für die Komplexität mittels deterministischer Verfahren,

$$n^{\mathrm{ran}}(\varepsilon, d, p) \;\le\; n^{\mathrm{det}}(\varepsilon, d, p) \;\le\; \exp\left(\mathrm{e}\, \log(d+1)\, \max\left\{\log \tfrac{1}{\varepsilon},\, d^{1/p}\right\}\right),$$

siehe Theorem 2.21. Die algorithmische Idee stammt von Vybíral [77], welcher ein Problem untersucht hat, das dem hiesigen Fall $p = \infty$ nahekommt.

Für dieses Beispiel beobachten wir grob gesprochen eine exponentielle Abhängigkeit der Komplexität von $d^{1/p}$. Dies kann durch Randomisierung nicht verbessert werden. Das betrachtete Problem ist zudem ein Beispiel dafür, wie eine Einschränkung der Input-Menge die Durchführbarkeit der Approximation beeinflusst.

### Zu Kapitel 3: Gleichmäßige Approximation von Funktionen aus einem Hilbert-Raum

Dieses Kapitel enthält einen neuen Monte-Carlo-Ansatz für die $L_\infty$-Approximation von Funktionen aus einem Hilbert-Raum $\mathcal{H}$ mit reproduzierendem Kern. Die Menge $F$ der Eingabegrößen ist die Einheitskugel in $\mathcal{H}$, d.h. für eine Orthonormalbasis $(\psi_k)_{k\in\mathbb{N}}$ von $\mathcal{H}$ haben wir

$$F := \left\{ \sum_{k=1}^{\infty} a_k\, \psi_k \;\middle|\; a_k \in \mathbb{R}, \sum_{k=1}^{\infty} a_k^2 \le 1 \right\}.$$

Die Idee für den neuen Algorithmus basiert auf einer fundamentalen Monte-Carlo-Approximationsmethode nach Mathé [46], siehe auch Abschnitt 3.2.1. Jene wurde

in der Originalarbeit zur Rekonstruktion in endlichdimensionalen Folgenräumen $\ell_2^m \hookrightarrow \ell_q^m$, $q > 2$, angewandt und diente in Verbindung mit Diskretisierungstechniken für Funktionenräume der Bestimmung der Konvergenzordnung für Einbettungsoperatoren. In Abschnitt 3.3.1 verfolgen wir einen direkteren Ansatz über die lineare Monte-Carlo-Methode

$$A_n^\omega(f) := \frac{1}{n} \sum_{i=1}^{n} L_i^\omega(f)\, g_i^\omega \,,$$

wobei

$$L_i^\omega(f) := \sum_{k=1}^{\infty} X_{ik}\, \langle \psi_k, f \rangle_{\mathcal{H}}, \quad \text{und} \quad g_i^\omega := \sum_{k=1}^{\infty} X_{ik}\, \psi_k \,,$$

mit unabhängigen standardnormalverteilten Zufallsvariablen $X_{ik}$. Die Funktionen $g_i^\omega$ sind unabhängige Realisierungen des mit $\mathcal{H}$ assoziierten *Gauß-Feldes* $\Psi$, die Kovarianzfunktion von $\Psi$ ist der reproduzierende Kern von $\mathcal{H}$. Für dieses Verfahren gilt die Fehlerabschätzung

$$e((A_n^\omega)_\omega, F) \leq \frac{2\, \mathbb{E}\, \|\Psi\|_\infty}{\sqrt{n}} \,.$$

Zugegebenermaßen sind die zufälligen Funktionale $L_i^\omega$ unstetig mit Wahrscheinlichkeit 1, für festes $f \in \mathcal{H}$ jedoch sind die Werte $L_i^\omega(f)$ zentrierte Gauß-Variablen mit Varianz $\|f\|_{\mathcal{H}}^2$ und somit fast sicher endlich. Das Verfahren $(A_n^\omega)_\omega$ kann allerdings auch als Grenzwert von Methoden gesehen werden, die fast sicher stetige Funktionale verwenden, siehe Lemma 3.3.

Mit Werkzeugen aus der Stochastik, siehe Abschnitt 3.3.3 für eine Zusammenstellung, können wir den Wert $\mathbb{E}\, \|\Psi\|_\infty$ abschätzen, sofern die zufällige Funktion $\Psi$ beschränkt ist. Insbesondere mit der Technik *majorisierender Maße* nach Fernique lässt sich der Fall periodischer Funktionen auf dem $d$-dimensionalen Torus $\mathbb{T}^d$ angehen, siehe Abschnitt 3.4.2. Hierbei ist $\mathbb{T}$ das Intervall $[0, 1]$ mit identifizierten Randpunkten. Im eindimensionalen Fall bezeichnen wir mit $\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T})$ den Raum mit Orthonormalbasis

$$\{\lambda_0,\ \lambda_k\, \sin(2\,\pi\, k\, \cdot),\ \lambda_k\, \cos(2\,\pi\, k\, \cdot)\}_{k \in \mathbb{N}} \,,$$

wobei $\lambda_k > 0$. Der $d$-variate Fall ist über das Tensorprodukt definiert,

$$\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d) := \bigotimes_{j=1}^{d} \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}) \,.$$

Wir nehmen $\sum_{k=0}^{\infty} \lambda_k^2 \stackrel{!}{=} 1$ an, sodass der Anfangsfehler konstant 1 ist. Unter diesen Voraussetzungen gilt der Fluch der Dimension für deterministische Verfahren, siehe Theorem 3.15. Die deterministische untere Schranke basiert auf einer Beweistechnik von Kuo, Wasilkowski und Woźniakowski [40], ebenso Cobos, Kühn und Sickel [12], siehe Abschnitt 3.3.4. Im randomisierten Fall leiten wir Bedingungen an den reproduzierenden Kern periodischer Hilbert-Räume ab, für die der assoziierte Gauß-Prozess beschränkt ist. Im Speziellen betrachten wir Korobov-Räume $H_r^{\mathrm{Kor}}(\mathbb{T}^d) = \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d)$ mit $\lambda_k := \sqrt{\beta_1}\, k^{-r}$ für $k \in \mathbb{N}$. Hierbei ist $\beta_1 > 0$ so gewählt, dass

der Anfangsfehler immer noch über die Wahl von $0 < \lambda_0 < 1$ angepasst werden kann. Für Glattheit $r > 1$ lässt sich zeigen, dass das Approximationsproblem

$$\mathrm{APP} : H_r^{\mathrm{Kor}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)$$

eine polynomiell beschränkte Monte-Carlo-Komplexität besitzt,

$$n^{\mathrm{ran}}(\varepsilon, d, r) \leq C_r \, d \, (1 + \log d) \, \varepsilon^{-2} \,,$$

wobei $C_r > 0$. Für weniger Glattheit, konkret $\frac{1}{2} < r \leq 1$, können wir immer noch Durchführbarkeit der Approximation mit polynomiell beschränktem Aufwand (*polynomial tractability*) zeigen, wobei die Schranken für die Komplexität schlechter werden. Hierbei wird die fundamentale Monte-Carlo-Methode nur noch auf endlichdimensionale Teilräume von $H_r^{\mathrm{Kor}}(\mathbb{T}^d)$ angewandt, siehe Theorem 3.19. Auf diese Weise bricht Monte Carlo den Fluch.

### Zu Kapitel 4: Approximation monotoner Funktionen

Wir untersuchen das Problem der $L_1$-Approximation für die Klasse beschränkter, monotoner Funktionen,

$$F_{\mathrm{mon}}^d := \{ f : [0,1]^d \to [0,1] \mid \mathbf{x} \leq \mathbf{z} \Rightarrow f(\mathbf{x}) \leq f(\mathbf{z}) \} \,,$$

unter Nutzung von Funktionswerten als Information. Dies ist kein lineares Problem, weil die Inputmenge asymmetrisch ist. Hinrichs, Novak und Woźniakowski [28] zeigten, dass das Problem im deterministischen Fall dem Fluch der Dimension unterliegt. Dies ist bei Randomisierung nicht mehr der Fall, dennoch bleibt das Problem sehr schwer zu lösen.

Aus einem Ergebnis von Blum, Burch und Langford [8] für Boole'sche monotone Funktionen $f : \{0,1\}^d \to \{0,1\}$ kann man ableiten, dass für festes $\varepsilon > 0$ die Komplexität $n^{\mathrm{ran}}(\varepsilon, d)$ mindestens exponentiell von $\sqrt{d}$ abhängt. Abschnitt 4.3 enthält einen modifizierten Beweis, dank dem wir eine untere Schranke mit aussagekräftiger $\varepsilon$-Abhängigkeit bekommen,

$$n^{\mathrm{ran}}(\varepsilon, d) > \nu \exp(c \sqrt{d} \, \varepsilon^{-1}) \quad \text{für } \varepsilon_0 \sqrt{d_0/d} \leq \varepsilon \leq \varepsilon_0 \,,$$

wobei $\varepsilon_0, \nu, c > 0$ und $d \geq d_0 \in \mathbb{N}$, siehe Theorem 4.8. Insbesondere wenn wir eine gemäßigt abfallende Folge von Fehlertoleranzen $\varepsilon_d := \varepsilon_0 \sqrt{d_0/d}$ wählen, lässt sich beobachten, dass die Komplexität $n^{\mathrm{ran}}(\varepsilon_d, d)$ exponentiell in $d$ wächst. Man sagt, das Problem sei *nicht „weakly tractable“*, siehe Remark 4.9.

In Abschnitt 4.4 werden obere Schranken bewiesen, die zeigen, dass die Komplexität $n^{\mathrm{ran}}(\varepsilon, d)$ für festes $\varepsilon > 0$ tatsächlich „nur“ exponentiell von $\sqrt{d}$ modulo logarithmischer Terme abhängt. Die algorithmische Idee wurde bereits von Bshouty und Tamon [9] für *Boole'sche* monotone Funktionen umgesetzt, siehe Abschnitt 4.4.1. Ein vergleichbarer Ansatz für *reellwertige*, auf $[0,1]^d$ definierte, monotone Funktionen wird nun in Abschnitt 4.4.2 verfolgt. Darin beschreiben und analysieren wir einen neuen Monte-Carlo-Algorithmus $(A_{r,k,n}^{\omega})_{\omega}$ mit wünschenswerten Fehlerschranken,

hierbei $r, k, n \in \mathbb{N}$. Im Wesentlichen basiert dieser auf einer Standard Monte-Carlo-Näherung für die wichtigsten Wavelet-Koeffizienten der Haar-Basis in $L_2([0,1]^d)$, wobei die zu approximierende Funktion an $n$ zufällig gewählten Stellen ausgewertet wird. Die ausgegebene Funktion ist konstant auf Teilwürfeln der Seitenlänge $2^{-r}$, d.h. nur Wavelet-Koeffizienten bis zu einer bestimmten Auflösung kommen in Betracht. Außerdem sind nur solche Wavelet-Koeffizienten von Interesse, die – für eine Input-Funktion $f$ – die gleichzeitige Abhängigkeit von bis zu $k$ Variablen messen. Für festes $\varepsilon$ hat dieser Parameter das asymptotische Verhalten $k \asymp \sqrt{d\,(1+\log d)}$. Es gibt eine lineare Version des Algorithmus, siehe Theorem 4.22, sowie eine nichtlineare mit verbesserter $\varepsilon$-Abhängigkeit der Komplexität, siehe Remark 4.23.

# Introduction and Results

For many problems arising in technical and scientific applications it is practically impossible to give exact solutions. Instead, one is interested in approximate solutions that are to be found with methods that perform a finite number of steps. In particular, we need to cope with incomplete information about a problem instance; apart from structural assumptions (the so-called *a priori* knowledge), we may collect only a finite amount of information, let us say $n$ real numbers originating from measurements or from subprograms provided by the user. There is a growing interest in solving high-dimensional problems that involve functions defined on a $d$-dimensional domain. We study the so-called *information-based complexity* $n(\varepsilon, d)$, that is the minimal number of information needed in order to solve the problem within a given error tolerance $\varepsilon > 0$. *Tractability* studies in general are concerned with the behaviour of this function $n(\varepsilon, d)$. In many cases the complexity increases exponentially in $d$ for some fixed $\varepsilon$, this phenomenon is called the *curse of dimensionality*. If a problem suffers from the curse of dimensionality, there are basically two ways to deal with it. One way is to include more a priori knowledge, thus narrowing the set of possible inputs. The other way is to widen the class of admissible algorithms. In this dissertation we focus on the second approach, namely, we study the potential of randomization for function approximation problems.

A $d$-dimensional function approximation problem is an embedding

$$\mathrm{APP} : F^d \hookrightarrow G^d, \quad f \mapsto f\,,$$

with an *input set* $F^d$ which contains $d$-variate real-valued functions, and a normed space $G^d$. In this study, functions from $F^d$ are usually defined on the $d$-dimensional unit cube $[0,1]^d$, the output space $G^d$ could be $L_1([0,1]^d)$ or $L_\infty([0,1]^d)$. Deterministic algorithms are mappings $A_n = \phi \circ N : F^d \to G^d$, where the *information mapping*

$$N : F^d \to \mathbb{R}^n, \quad f \mapsto (L_1(f), \ldots, L_n(f))\,,$$

uses $n$ linear functionals $L_i$ as information about the problem instance $f$, and $\phi : \mathbb{R}^n \to G^d$. The error is then defined by the worst case,

$$e(A_n, F^d) := \sup_{f \in F^d} \|f - A_n(f)\|_{G^d}\,.$$

Randomized methods are modelled as a family $(A_n^\omega)_{\omega\in\Omega}$ of mappings $A_n^\omega : F^d \to G^d$ as before, where $\omega \in \Omega$ is a random element from a probability space $(\Omega, \Sigma, \mathbb{P})$. The

error of such a *Monte Carlo* algorithm is defined as the expected error for the worst input,

$$e((A_n^\omega)_\omega, F^d) := \sup_{f \in F^d} \mathbb{E}\, \|f - A_n^\omega(f)\|_{G^d}\,.$$

In both cases, the aim is to significantly reduce the *initial error*

$$e(0, F^d) := \inf_{g \in G^d} \sup_{f \in F^d} \|f - g\|_{G^d}\,,$$

which is achievable without any information. We are interested in the comparison of the complexity in the deterministic and the randomized setting,

$$n^{\mathrm{det}}(\varepsilon, d) := \inf\{n \in \mathbb{N}_0 \mid \exists A_n : e(A_n, F^d) \le \varepsilon\}\,,$$
$$n^{\mathrm{ran}}(\varepsilon, d) := \inf\{n \in \mathbb{N}_0 \mid \exists (A_n^\omega)_\omega : e((A_n^\omega)_\omega, F^d) \le \varepsilon\}\,,$$

where $0 < \varepsilon < e(0, F^d)$. All algorithms presented in this thesis for upper bounds on these quantities are *non-adaptive* algorithms with the simple structure of $N$ as indicated above. The lower bounds are proven for more general *adaptive* algorithms, even *varying cardinality* $n(\omega, f)$ is considered. For these notions and a detailed introduction to *information-based complexity* see Chapter 1.

New results are contained in Chapters 2–4, which treat more or less stand-alone topics. Chapter 2 is concerned with lower bounds for randomized methods, by means of which in some cases one can show that Monte Carlo methods are not much better than optimal deterministic algorithms. In contrast to this, in Chapter 3 we find settings where deterministic algorithms suffer from the curse of dimensionality but randomization can break the curse in a very impressive way. Chapter 4 deals with a problem for which randomization breaks the curse of dimensionality, yet the problem is quite difficult.

### On Chapter 2: Lower Bounds for Linear Problems via Bernstein Numbers

The main result of this chapter is a lower bound for Monte Carlo algorithms for general linear problems

$$S : F \to G\,,$$

that is, $S$ is a linear operator between normed spaces $\widetilde{F}$ and $G$, and the input set $F$ is the unit ball in $\widetilde{F}$. We show that for any adaptive Monte Carlo method $(A_n^\omega)_\omega$ using $n$ arbitrary continuous linear functionals $L_i$ as information, we have

$$e((A_n^\omega)_\omega, F) \ge \tfrac{1}{30}\, b_{2n}(S)\,,$$

where $b_m(S)$ is the $m$-th Bernstein number of the operator $S$, see Theorem 2.1. The proof is based on a result due to Heinrich [22] which connects norm expectations for Gaussian measures with the Monte Carlo error. The innovation is that we use Lewis' theorem for choosing optimal Gaussian measures. This result has been announced in [39], a short proof without the explicit constant has been included there.

In Section 2.4.2 we apply this general tool to the $L_\infty$-approximation of certain classes of $C^\infty$-functions,

$$\mathrm{APP} : F_p^d \hookrightarrow L_\infty([0,1]^d)\,.$$

Here, the input set is defined as

$$F_p^d := \{f \in C^\infty([0,1]^d) \mid \|\nabla_{\mathbf{v}_k} \cdots \nabla_{\mathbf{v}_1} f\|_\infty \le |\mathbf{v}_1|_p \cdots |\mathbf{v}_k|_p \\ \text{for } k \in \mathbb{N}_0,\, \mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^d\},$$

where $\nabla_{\mathbf{v}} f$ denotes the directional derivative along a vector $\mathbf{v} \in \mathbb{R}^d$, and we write $|\mathbf{v}|_p$ for the $p$-norm of $\mathbf{v} \in \ell_p^d$, $1 \le p \le \infty$. Via the corresponding Bernstein numbers we obtain the lower bound

$$n^{\mathrm{ran}}(\varepsilon, d, p) > 2^{\lfloor \frac{d^{1/p}}{3} \rfloor - 1} \quad \text{for } 0 < \varepsilon \le \tfrac{1}{30},$$

see Corollary 2.20. For $p = 1$ this implies the curse of dimensionality even in the randomized setting. The technique for determining the Bernstein numbers is known from Novak and Woźniakowski [56], where the curse of dimensionality for the case $p = 1$ was shown in the deterministic setting.

A simple Taylor approximation provides upper bounds for the complexity with deterministic methods,

$$n^{\mathrm{ran}}(\varepsilon, d, p) \le n^{\mathrm{det}}(\varepsilon, d, p) \le \exp\left(\mathrm{e}\, \log(d+1)\, \max\left\{\log \tfrac{1}{\varepsilon},\, d^{1/p}\right\}\right),$$

see Theorem 2.21. The algorithmic idea goes back to Vybíral [77] who considered a setting similar to the case $p = \infty$ here.

For this example we observe an exponential dependency of the complexity on $d^{1/p}$, roughly, which cannot be removed with randomization. It is also an example which shows how narrowing the input set may affect tractability.

### On Chapter 3: Uniform Approximation of Functions from a Hilbert Space

In this chapter we study a new Monte Carlo approach to the $L_\infty$-approximation of functions from a reproducing kernel Hilbert space $\mathcal{H}$. The input set $F$ is the unit ball of $\mathcal{H}$, that is, for an orthonormal basis $(\psi_k)_{k \in \mathbb{N}}$ of $\mathcal{H}$ we have

$$F := \left\{ \sum_{k=1}^\infty a_k\, \psi_k \;\middle|\; a_k \in \mathbb{R}, \sum_{k=1}^\infty a_k^2 \le 1 \right\}.$$

The idea for the new algorithm is based on a fundamental Monte Carlo approximation method which is due to Mathé [46], see also Section 3.2.1. In the original paper it has been applied to finite dimensional sequence recovery $\ell_2^m \hookrightarrow \ell_q^m$, $q > 2$, it was then used in combination with discretization techniques for function space embeddings in order to determine the order of convergence. In Section 3.3.1 we take a more direct approach, proposing the linear Monte Carlo method

$$A_n^\omega(f) := \frac{1}{n} \sum_{i=1}^n L_i^\omega(f)\, g_i^\omega$$

where

$$L_i^\omega(f) := \sum_{k=1}^\infty X_{ik}\, \langle \psi_k, f \rangle_{\mathcal{H}}, \quad \text{and} \quad g_i^\omega := \sum_{k=1}^\infty X_{ik}\, \psi_k,$$

with the $X_{ik}$ being independent standard Gaussian random variables. The functions $g_i^\omega$ are independent copies of the *Gaussian field* $\Psi$ associated to $\mathcal{H}$, the covariance function of $\Psi$ is the reproducing kernel of $\mathcal{H}$. We have the error estimate

$$e((A_n^\omega)_\omega, F) \leq \frac{2\,\mathbb{E}\,\|\Psi\|_\infty}{\sqrt{n}}\,.$$

Admittedly, the random functionals $L_i^\omega$ are discontinuous with probability 1, but for any fixed $f \in \mathcal{H}$ the value $L_i^\omega(f)$ is a zero-mean Gaussian random variable with variance $\|f\|_{\mathcal{H}}^2$, hence it is almost surely finite. This method, however, can be seen as the limiting case of methods that use continuous random functionals, see Lemma 3.3.

Using tools from stochastics, see Section 3.3.3 for a summary, we can estimate the value $\mathbb{E}\,\|\Psi\|_\infty$, provided that the random function $\Psi$ is bounded. Namely, via the technique of *majorizing measures* due to Fernique, we tackle the case of periodic functions on the $d$-dimensional torus $\mathbb{T}^d$, see Section 3.4.2. Here, $\mathbb{T}$ is the interval $[0,1]$ where the endpoints are identified. In the univariate case, we denote by $\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T})$ the space with orthonormal basis

$$\{\lambda_0,\ \lambda_k\,\sin(2\,\pi\,k\,\cdot),\ \lambda_k\,\cos(2\,\pi\,k\,\cdot)\}_{k\in\mathbb{N}}\ ,$$

where $\lambda_k > 0$. The $d$-variate case is defined by the tensor product,

$$\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d) := \bigotimes_{j=1}^{d} \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T})\,.$$

We assume $\sum_{k=0}^{\infty} \lambda_k^2 \stackrel{!}{=} 1$, and then the initial error is constant 1. For this situation we obtain the curse of dimensionality in the deterministic setting, see Theorem 3.15. The deterministic lower bound is based on a technique due to Kuo, Wasilkowski, and Woźniakowski [40], also Cobos, Kühn, and Sickel [12], see Section 3.3.4. For the randomized setting, we derive conditions on the reproducing kernel of periodic Hilbert spaces such that the associated Gaussian process is bounded. Specifically, we consider Korobov spaces $H_r^{\mathrm{Kor}}(\mathbb{T}^d) = \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d)$ with $\lambda_k := \sqrt{\beta_1}\,k^{-r}$ for $k \in \mathbb{N}$, here $\beta_1 > 0$ such that the initial error may still be adjusted with $0 < \lambda_0 < 1$. For smoothness $r > 1$ we can show that the approximation problem

$$\mathrm{APP} : H_r^{\mathrm{Kor}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)$$

possesses a polynomially bounded Monte Carlo complexity,

$$n^{\mathrm{ran}}(\varepsilon, d, r) \leq C_r\,d\,(1+\log d)\,\varepsilon^{-2}\,,$$

where $C_r > 0$. Hence this problem is *polynomially tractable*. For smaller smoothness $\frac{1}{2} < r \leq 1$, we can still prove polynomial tractability with a worse complexity bound, in that case the fundamental Monte Carlo method is only applied to a finite dimensional subspace of $H_r^{\mathrm{Kor}}(\mathbb{T}^d)$, see Theorem 3.19. By this, Monte Carlo breaks the curse.

### On Chapter 4: Approximation of Monotone Functions

We study the $L_1$-approximation for the class of bounded monotone functions,

$$F_{\mathrm{mon}}^d := \{f : [0,1]^d \to [0,1] \mid \mathbf{x} \le \mathbf{z} \Rightarrow f(\mathbf{x}) \le f(\mathbf{z})\},$$

based on *function values* as information. This problem is not linear since the input set is unbalanced. Hinrichs, Novak, and Woźniakowski [28] showed that the problem suffers from the curse of dimensionality in the deterministic setting. This is not the case in the randomized setting anymore, still the problem is very difficult.

From a result by Blum, Burch, and Langford [8] for monotone Boolean functions $f : \{0,1\}^d \to \{0,1\}$, we can conclude that for fixed $\varepsilon > 0$ the complexity $n^{\mathrm{ran}}(\varepsilon, d)$ depends exponentially on $\sqrt{d}$ at least. In Section 4.3 a modified proof is given by what we obtain a lower bound that includes a meaningful $\varepsilon$-dependency,

$$n^{\mathrm{ran}}(\varepsilon, d) > \nu \exp(c\sqrt{d}\,\varepsilon^{-1}) \quad \text{for } \varepsilon_0\sqrt{d_0/d} \le \varepsilon \le \varepsilon_0,$$

where $\varepsilon_0, \nu, c > 0$ and $d \ge d_0 \in \mathbb{N}$, see Theorem 4.8. In particular, choosing a moderately decaying sequence of error tolerances $\varepsilon_d := \varepsilon_0\sqrt{d_0/d}$, we observe that the complexity $n^{\mathrm{ran}}(\varepsilon_d, d)$ grows exponentially in $d$. This implies that the problem is *not weakly tractable*, see Remark 4.9.

In Section 4.4 we prove upper bounds which show that, for fixed $\varepsilon > 0$, the complexity $n^{\mathrm{ran}}(\varepsilon, d)$ indeed depends exponentially on $\sqrt{d}$ times some logarithmic terms only. The algorithmic idea has been performed for monotone Boolean functions in Bshouty and Tamon [9], see Section 4.4.1. Inspired by this, in Section 4.4.2 a new Monte Carlo algorithm $(A^{\omega}_{r,k,n})_{\omega}$ with desirable error bounds for real-valued monotone functions defined on $[0,1]^d$ is proposed and studied, here $r, k, n \in \mathbb{N}$. Essentially, we use standard Monte Carlo approximation for the most important wavelet coefficients of the Haar basis in $L_2([0,1]^d)$, using $n$ random samples. The output will be constant on subcubes of sidelength $2^{-r}$, so only wavelet coefficients up to a certain resolution come into consideration. Further, only those wavelet coefficients are of interest that – for an input function $f$ – measure the simultanious dependency on at most $k$ variables. For fixed $\varepsilon$, this parameter has the asymptotic behaviour $k \asymp \sqrt{d\,(1+\log d)}$. There is a linear version of the algorithm, see Theorem 4.22, and a non-linear version with improved $\varepsilon$-dependency of the complexity, see Remark 4.23.

# Contents

**Zusammenfassung** **i**

**Introduction and Results** **vii**

**1 Basic Notions in Information-Based Complexity** **1**
1.1 Types of Errors and Information . . . . . 1
1.2 Tractability . . . . . 6
1.3 Algorithms with Varying Cardinality . . . . . 9
1.4 On the Measurability of Algorithms . . . . . 13

**2 Lower Bounds for Linear Problems via Bernstein Numbers** **16**
2.1 The Setting and Bernstein Numbers . . . . . 16
2.2 Adaptive Monte Carlo Methods . . . . . 18
2.2.1 The Conditional Measure . . . . . 20
2.2.2 Norm Expectation of Truncated Gaussian Measures . . . . . 24
2.2.3 Optimal Gaussian Measures . . . . . 26
2.3 Special Settings . . . . . 31
2.3.1 Varying Cardinality . . . . . 31
2.3.2 Homogeneous Monte Carlo Methods . . . . . 33
2.4 Applications . . . . . 36
2.4.1 Recovery of Sequences . . . . . 36
2.4.2 Approximation of Ultra Smooth Functions . . . . . 41

**3 Uniform Approximation of Functions from a Hilbert Space** **45**
3.1 Motivation and the General Setting . . . . . 45
3.2 Introduction to Randomized Approximation . . . . . 46
3.2.1 A Fundamental Monte Carlo Approximation Method . . . . . 47
3.2.2 Methods for the Recovery of Sequences . . . . . 49
3.2.3 Speeding up the Convergence for Function Approximation . . 53
3.2.4 Breaking the Curse - a Sequence Space Modell . . . . . 56
3.3 Tools for Function Approximation . . . . . 57
3.3.1 A Plain Monte Carlo Upper Bound . . . . . 58
3.3.2 Reproducing Kernel Hilbert Spaces . . . . . 60
3.3.3 Expected Maximum of Zero-Mean Gaussian Fields . . . . . 63
3.3.4 A Lower Bound in the Worst Case Setting . . . . . 66
3.4 Breaking the Curse for Function Approximation . . . . . 72

3.4.1 Approximation with the Brownian Sheet . . . . . 72
3.4.2 Tensor Product Spaces of Periodic Functions . . . . . 76
3.4.3 Final Remarks on the Initial Error . . . . . 87

**4 Approximation of Monotone Functions 89**
4.1 The Setting and Background . . . . . 89
4.2 First Simple Estimates . . . . . 91
4.2.1 The Classical Approach – Order of Convergence . . . . . 92
4.2.2 Curse of Dimensionality in the Deterministic Setting . . . . . 95
4.3 Intractability for Randomized Approximation . . . . . 98
4.3.1 The Result – A Monte Carlo Lower Bound . . . . . 98
4.3.2 The Proof of the Monte Carlo Lower Bound . . . . . 102
4.3.3 Remarks on the Proof . . . . . 110
4.4 Breaking the Curse with Monte Carlo . . . . . 114
4.4.1 A Known Method for Boolean Functions . . . . . 114
4.4.2 Real-Valued Monotone Functions . . . . . 119

**A On Gaussian Measures 128**
A.1 Comparison of Gaussian Measures . . . . . 128
A.2 Restrictions of Gaussian Measures . . . . . 129
A.3 Deviation Estimates for Gaussian Measures . . . . . 129
A.4 Gaussian Vectors in Sequence Spaces . . . . . 130

**B Useful Inequalities 134**
B.1 Combinatorics . . . . . 134
B.2 Quantitative Central Limit Theorem . . . . . 135

**Abbreviations and Symbols 136**

**Bibliography 138**

# Chapter 1

# Basic Notions in Information-Based Complexity

In Section 1.1 the basic notions for the model of computation and approximation in *information-based complexity* (IBC) are introduced. In Section 1.2 on *tractability* we provide the notions for a classification of multi-dimensional problems by the difficulty of solving them. After these two sections the reader may immediately go forward to one of the three main chapters (Chapters 2–4) that cover different topics. Section 1.3 on algorithms with varying cardinality is an extension of the computational model, we collect tools that help to extend lower bounds to this broader class of algorithms. Section 1.4 is a comment on the computational model, especially on measurability assumptions, it has no further connection to the rest of the thesis.

## 1.1 Types of Errors and Information

We collect all notions we need for a basic understanding of *information-based complexity* (IBC). For an elaborate introduction to this field, we refer to the book of Traub, Wasilkowski, and Woźniakowski [73].

Let $S : \widetilde{F} \to G$ be the so-called *solution mapping* between the *input space* $\widetilde{F}$, and the *target space* $G$ which is a metric space. We aim to approximate $S$ for inputs from an *input set* $F \subseteq \widetilde{F}$ with respect to the metric $\mathrm{dist}_G$ of the *target space* $G$, using algorithms that collect only a limited amount of information on the input $f \in F$ by evaluating finitely many functionals from a given class $\Lambda$.

A very common example are *linear problems* for which

- $S$ is a linear operator between Banach spaces,
- the input set $F$ is the unit ball in $\widetilde{F}$, or – more generally – a centrally symmetric convex set, and
- the class $\Lambda$ of all admissible functionals is a subclass of the class $\Lambda^{\mathrm{all}}$ of *all* continuous linear functionals.

Chapters 2 and 3 deal with linear problems. In Chapter 4, however, we will consider an input set $F$ consisting of monotone functions which is not centrally symmetric. Within

this research we mainly examine approximation problems $S = \mathrm{APP} : \widetilde{F} \hookrightarrow G$, $f \mapsto f$, that is, $\widetilde{F}$ is identified with a subset of $G$. Another typical example for problems is the computation of the definite integral $S = \mathrm{INT} : \widetilde{F} \to \mathbb{R}$, $f \mapsto \int_{[0,1]^d} f(\mathbf{x})\,\mathrm{d}\mathbf{x}$, with $\widetilde{F}$ being a class of integrable functions $f : [0,1]^d \to \mathbb{R}$; here, algorithms may use function values, also called *standard information* $\Lambda^{\mathrm{std}}$.

Let $(\Omega, \Sigma, \mathbb{P})$ be a suitable probability space. Further, let $\mathcal{B}(G)$ denote the Borel $\sigma$-algebra of $G$, and $\mathcal{F}$ be a suitable $\sigma$-algebra on $\widetilde{F}$, e.g. the Borel $\sigma$-algebra if $\widetilde{F}$ is a metric space. By *randomized algorithms*, also called *Monte Carlo algorithms*, we understand $(\Sigma \otimes \mathcal{F}) - \mathcal{B}(G)$-measurable mappings $A_n = (A_n^\omega(\cdot))_{\omega\in\Omega} : \Omega \times \widetilde{F} \to G$. This means that the output $A_n(f)$ for an input $f$ is random, depending on $\omega \in \Omega$. We consider algorithms of *cardinality* $n$ that use at most $n$ pieces of information,[1] i.e. $A_n^\omega = \phi^\omega \circ N^\omega$ where $N^\omega : \widetilde{F} \to \mathbb{R}^n$ is the so-called *information mapping*. The mapping $\phi^\omega : \mathbb{R}^n \to G$ generates an output $g = \phi^\omega(\mathbf{y}) \in G$ as a compromise for all possible inputs that lead to the same information $\mathbf{y} = N^\omega(f) \in \mathbb{R}^n$.[2] If, for any information vector $\mathbf{y} \in N^\omega(F)$, we take the output $\phi^\omega(\mathbf{y}) = S(\tilde{f})$ as the solution for an element $\tilde{f} \in F$ from the input set which *interpolates* the data, that means $N^\omega(\tilde{f}) = \mathbf{y}$, then the algorithm is called *interpolatory*. The combinatory cost for the computation of $\phi$ (arithmetic operations, comparison of real numbers, operations in $G$) is usually neglected.[3]

There are different types of information mappings. In this research the information is obtained by computing $n$ *functionals* from the class $\Lambda$ for the particular input. This could be function values $\Lambda^{\mathrm{std}}$, or arbitrary continuous linear functionals $\Lambda^{\mathrm{all}}$. We do not care about how these functionals are evaluated – they could be obtained by some measuring device or by a subroutine provided by the user – to us, evaluating an information functional is an *oracle call*. An information mapping is called *non-adaptive*, if

$$N^\omega(f) = [L_1^\omega(f), \ldots, L_n^\omega(f)] = (y_1, \ldots, y_n) = \mathbf{y}\,, \tag{1.1.1}$$

where all functionals $L_k^\omega \in \Lambda$ are chosen independently from $f$. In that case, $N^\omega$ is a linear mapping for fixed $\omega \in \Omega$. For *adaptive* information $N^\omega$ the choice of the functionals may depend on previously obtained information, we assume that the choice of the $k$-th functional is a measurable mapping $(\omega; \mathbf{y}_{[k-1]}) \mapsto L^\omega_{k;\mathbf{y}_{[k-1]}}(\cdot)$ into the space of functionals, here, $\mathbf{y}_{[k]} = (y_1, \ldots, y_k)$ for $k = 1, \ldots, n$. Further, the mapping $N = (N^\omega)_\omega : \Omega \times \widetilde{F} \to \mathbb{R}^n$ as a whole shall be $(\Sigma \otimes \mathcal{F}) - \mathcal{B}(\mathbb{R}^n)$-measurable. By $\mathcal{A}_n^{\mathrm{ran,ada}}(\Lambda)$ we denote the class of all Monte Carlo algorithms that use $n$ pieces of adaptively obtained information, for the subclass of non-adaptive algorithms we write $\mathcal{A}_n^{\mathrm{ran,nonada}}(\Lambda)$.

If the solution operator $S$ is a linear operator that maps between Banach spaces, we consider two more special types of algorithms.
*Linear* algorithms $\mathcal{A}_n^{\mathrm{ran,lin}}(\Lambda)$ comprise non-adaptive algorithms where not only $N^\omega$,

[1] See Section 1.3 for the extension of the computational model to algorithms with varying cardinality.

[2] Some authors call $\phi^\omega$ an *algorithm* and $\phi^\omega \circ N^\omega$ a *method*. In this dissertation, "method" and "algorithm" are used synonymously, both referring to $A_n^\omega = \phi^\omega \circ N^\omega$.

[3] We make one exception in Remark 4.24, where we compare two different outputs $\phi$.

but also $\phi^\omega$, and therefore $A_n^\omega = \phi^\omega \circ N^\omega$, is linear for every $\omega \in \Omega$. For linear algorithms we usually say *rank* instead of cardinality.
As another special class we consider *homogeneous* algorithms $\mathcal{A}_n^{\mathrm{ran,hom}}(\Lambda)$. The information mapping may still be adaptive, however with the special constraint

$$L^\omega_{k,\mathbf{y}_{[k-1]}} = L^\omega_{k,\lambda\,\mathbf{y}_{[k-1]}}$$

for information vectors $\mathbf{y} = N^\omega(f)$ and $\lambda \in \mathbb{R} \setminus \{0\}$. In particular, this implies homogeneity for the info mapping, $N^\omega(\lambda\, f) = \lambda\, N^\omega(f)$ for all $\lambda \in \mathbb{R}$ and $f \in \widetilde{F}$. For the mapping $\phi$ we assume the same, $\phi^\omega(\lambda\,\mathbf{y}) = \lambda\,\phi^\omega(\mathbf{y})$, thus inducing $A_n^\omega(\lambda\, f) = \lambda\, A_n^\omega(f)$.

We regard the class of *deterministic algorithms* as a subclass $\mathcal{A}_n^{\mathrm{det},\star} \subset \mathcal{A}_n^{\mathrm{ran},\star}$ ($\star \in \{\mathrm{ada}, \mathrm{nonada}, \mathrm{lin}, \mathrm{hom}\}$) of algorithms that are independent from $\omega \in \Omega$,[4] for a particular algorithm we write $A_n = \phi \circ N$, omitting the random element $\omega$. For a deterministic algorithm $A_n$, the (*absolute*) *error at* $f$ is defined as the distance between output and exact solution,

$$e(A_n, S, f) := \mathrm{dist}_G(A_n(f), S(f))\,. \tag{1.1.2}$$

For randomized algorithms $A_n = (A_n^\omega(\cdot))_{\omega\in\Omega}$, this can be generalized as the *expected error at* $f$,

$$e(A_n, S, f) := \mathbb{E}\,\mathrm{dist}_G(A_n^\omega(f), S(f))\,, \tag{1.1.3}$$

however, some authors prefer the *root mean square error*

$$e_2(A_n, S, f) := \sqrt{\mathbb{E}\,\mathrm{dist}_G(A_n^\omega(f), S(f))^2}\,. \tag{1.1.4}$$

(The expectation $\mathbb{E}$ is written for the integration over all $\omega \in \Omega$ with respect to $\mathbb{P}$.) Note that $e(A_n, S, f) \le e_2(A_n, S, f)$. Another criterion for rating Monte Carlo methods is the *margin of error*[5] for some preferably small *uncertainty level* $\delta \in (0, 1)$,

$$e_\delta(A_n, S, f) := \inf\{\varepsilon > 0 \mid \mathbb{P}(\mathrm{dist}_G(A_n^\omega(f), S(f)) > \varepsilon) \le \delta\}\,,$$

in other words, we have a confidence level $(1 - \delta)$ for the error $\varepsilon$. This criterion is more difficult to analyse than the other two definitions of a Monte Carlo error, however, a basic understanding of the power of randomization can already be gained with a simple mean error criterion.[6]
If the input space $\widetilde{F}$ is a normed space, one can also consider the *normalized* error criterion where for deterministic algorithms the error at $f \neq 0$ is defined as

$$e_{\mathrm{normal}}(A_n, S, f) := \frac{\|S(f) - A_n(f)\|_G}{\|f\|_F}\,. \tag{1.1.5}$$

The normalized error for randomized algorithms is defined analogously.

[4] This means in particular that we assume deterministic algorithms to be measurable. For a deeper discussion on measurability see Section 1.4.

[5] This is a common notion in statistics.

[6] In Section 4.4.1 we cite an algorithm proposed by Bshouty and Tamon [9]. They studied the margin of error, but we only reproduce the analysis for the expected error.

The *global error* of an algorithm $A_n$ is defined as the error for the worst input from the input set $F \subset \widetilde{F}$, we write

$$e(A_n, S, F) := \sup_{f \in F} e(A_n, S, f)\,. \tag{1.1.6}$$

For technical purposes, we also need the *$\mu$-average error*, which is defined for any (sub-)probability measure $\mu$ (the so-called *input distribution*) on the input space $\widetilde{F}$,

$$e(A_n, S, \mu) := \int e(A_n, S, f)\,\mathrm{d}\mu(f)\,. \tag{1.1.7}$$

(A sub-probability measure $\mu$ on $\widetilde{F}$ is a positive measure with $0 < \mu(\widetilde{F}) \leq 1$.)

The difficulty of a problem within a particular setting refers to the error of optimal algorithms, we define the *$n$-th minimal error*

$$\begin{aligned} e^{\diamond,\star}(n, S, F, \Lambda) &:= \inf_{A_n \in \mathcal{A}_n^{\diamond,\star}(\Lambda)} e(A_n, S, F) \quad \text{and} \\ e^{\diamond,\star}(n, S, \mu, \Lambda) &:= \inf_{A_n \in \mathcal{A}_n^{\diamond,\star}(\Lambda)} e(A_n, S, \mu)\,, \end{aligned}$$

where $\diamond \in \{\mathrm{ran}, \mathrm{det}\}$ and $\star \in \{\mathrm{ada}, \mathrm{nonada}, \mathrm{lin}, \mathrm{hom}\}$. These quantities are inherent properties of the problem $S$ with proper names. So, given an input set $F$, the worst input error for optimal randomized algorithms $e^{\mathrm{ran},\star}(n, S, F, \Lambda)$ is called the *Monte Carlo error*, the worst input error for deterministic algorithms $e^{\mathrm{det},\star}(n, S, F, \Lambda)$ is called the *worst case error* of the problem $S$. Given an input distribution $\mu$, we only consider deterministic algorithms, and – for better distinction from the other two settings – we introduce a new labelling $e^{\mathrm{avg},\star}(n, S, \mu, \Lambda) := e^{\mathrm{det},\star}(n, S, \mu, \Lambda)$, calling it the *$\mu$-average (case) error* of the problem $S$. For $n = 0$ we obtain the *initial error*, that is the minimal error that we achieve if we have to generate an output without collecting any information about the actual input.
The inverse notion is the *$\varepsilon$-complexity*[7] for a given error tolerance $\varepsilon > 0$,

$$n^{\diamond,\star}(\varepsilon, S, \bullet, \Lambda) := \inf\{n \in \mathbb{N}_0 \mid \exists_{A_n \in \mathcal{A}_n^{\diamond,\star}(\Lambda)},\, e(A_n, S, \bullet) \leq \varepsilon\}\,,$$

where $\bullet$ either stands for an input set $F \subset \widetilde{F}$, or for an input distribution $\mu$.

**Remark 1.1** (Monotonic properties of error quantities)**.** Obviously, in any setting, the error $e^{\diamond,\star}(n, S, \bullet, \Lambda)$ is monotonously decreasing (or steady) for growing $n$. Similarly, the inverse notion of complexity $n^{\diamond,\star}(\varepsilon, S, \bullet, \Lambda)$ is growing (or steady) for $\varepsilon \to 0$.

By definition, the error (or the complexity, respectively) is smaller or equal for smaller input sets $F' \subseteq F$,

$$e^{\diamond,\star}(n, S, F', \Lambda) \leq e^{\diamond,\star}(n, S, F, \Lambda)\,.$$

[7] More precisely, we should call this quantity *$\varepsilon$-information complexity*. In the book on IBC by Traub et al. [73] it is called *$\varepsilon$-cardinality*, whereas the notion *complexity* is associated to the total computational cost taking combinatory operations such as addition, multiplication, comparisons and evaluation of certain elementary functions into account.

In general, a broader class of algorithms $\mathcal{A}'(\Lambda) \supseteq \mathcal{A}(\Lambda)$ can only lead to a smaller error (and complexity), so, since adaption and randomization are additional features for algorithms, we have

$$e^{\mathrm{ran},\star}(n, S, \bullet, \Lambda) \le e^{\mathrm{det},\star}(n, S, \bullet, \Lambda) \quad \text{and} \quad e^{\diamond,\mathrm{ada}}(n, S, \bullet, \Lambda) \le e^{\diamond,\mathrm{nonada}}(n, S, \bullet, \Lambda)\,. \tag{1.1.8}$$

For the same reason, more general classes of information functionals $\Lambda' \supseteq \Lambda$ will diminish the error (and the complexity),

$$e^{\diamond,\star}(n, S, \bullet, \Lambda') \le e^{\diamond,\star}(n, S, \bullet, \Lambda)\,.$$

If for a particular problem function evaluations are continuous, then arbitrary continuous functionals are a generalization, so in that case we have $\Lambda^{\mathrm{all}} \supseteq \Lambda^{\mathrm{std}}$.

Another important relationship connects average errors and the Monte Carlo error. It has already been used by Bakhvalov [6, Sec 1].

**Proposition 1.2** (Bakhvalov's technique)**.** *Let $\mu$ be an arbitrary (sub-)probability measure supported on the input set $F \subseteq \widetilde{F}$. Then*

$$e^{\mathrm{ran},\star}(n, S, F, \Lambda) \ge e^{\mathrm{avg},\star}(n, S, \mu, \Lambda)\,.$$

*Proof.* Let $A_n = (A_n^\omega)_{\omega\in\Omega} \in \mathcal{A}_n^{\mathrm{ran},\star}(\Lambda)$ be a Monte Carlo algorithm. We find

$$\begin{aligned}
e(A_n, S, F) &= \sup_{f\in F} \mathbb{E}\, e(A_n^\omega, S, f) \\
&\ge \int \mathbb{E}\, e(A_n^\omega, S, f)\, \mu(\mathrm{d}f) \\
[\text{Fubini}] \quad &= \mathbb{E} \int e(A_n^\omega, S, f)\, \mu(\mathrm{d}f) \\
&= \mathbb{E}\, e(A_n^\omega, S, \mu) \\
&\ge \inf_\omega e(A_n^\omega, S, \mu) \\
&\ge \inf_{A_n' \in \mathcal{A}_n^{\mathrm{det},\star}(\Lambda)} e(A_n', S, \mu)\,.
\end{aligned}$$

Here, we used that for any fixed elementary event $\omega \in \Omega$ the realization $A_n^\omega$ can be seen as a deterministic algorithm. □

The proof of the above relation also shows

$$e^{\mathrm{ran},\star}(n, S, \mu, \Lambda) = e^{\mathrm{det},\star}(n, S, \mu, \Lambda) \equiv e^{\mathrm{avg},\star}(n, S, \mu, \Lambda)\,,$$

so there is no need for randomized algorithms in an average case setting.

**Remark 1.3** (On upper and lower bounds)**.** Bakhvalov's technique provides the standard tool for proving lower bounds for the Monte Carlo error by considering particular average case situations. This has the advantage that we have to deal only

with deterministic algorithms. We have freedom to choose a suitable distribution $\mu$.[8] The proof of upper bounds basically relies on the analysis of proposed algorithms. Mathé [47] showed that in several cases one can theoretically find input distributions $\mu$ supported on $F$ such that the $\mu$-average error matches the Monte Carlo error.

Lower (upper) bounds for the $n$-th minimal error correspond to lower (upper) bounds for the $\varepsilon$-complexity, in detail,

$$e^{\diamond,\star}(n_0, S, \bullet, \Lambda) > \varepsilon_0 \quad \Rightarrow \quad n^{\diamond,\star}(\varepsilon_0, S, \bullet, \Lambda) > n_0 \,.$$

Consequently, Bakhvalov's technique can also be written down in the notion of $\varepsilon$-complexity,

$$n^{\mathrm{ran},\star}(\varepsilon, S, F, \Lambda) \geq n^{\mathrm{avg},\star}(\varepsilon, S, \mu, \Lambda) \,.$$

If no confusion is possible, in the future we will use a reduced notation, e.g. writing $e^{\mathrm{ran}}(n, S)$ instead of $e^{\mathrm{ran,ada}}(n, S, F, \Lambda)$ if the class $\Lambda$ of information functionals is known from the context, the input set $F$ is the unit ball of the input space $\widetilde{F}$ in the setting of a linear operator $S$ between Banach spaces, and taking into account that adaptive algorithms are the most general type of algorithms we consider. The same applies for the complexity $n^{\mathrm{ran}}(\varepsilon, S)$. In any case, the notation should be compact, yet include all aspects needed to distinguish different settings within the context.

## 1.2 Tractability

We give a short overview over different notions used in tractability theory. For a more detailed introduction refer to the book by Novak and Woźniakowski [55, Chap 2].

In tractability analysis we do not just consider a single solution operator but an entire family of solution operators

$$(S^d : F^d \to G^d)_{d \in \mathbb{N}},$$

with $d$ being a dimensional parameter. This could mean, for example, that $F^d$ and $G^d$ are classes of $d$-variate functions defined on the unit cube $[0,1]^d$.

In classical numerical analysis, however, the dimension $d$ is typically considered a fixed parameter – along with smoothness parameters etc. – so within a complexity setting[9] the error is perceived as a function of $n$,

$$e^{\diamond,\star}(n, S^d, F^d, \Lambda) = e(n) \,.$$

For solvable problems this function is non-increasing and converging to 0 for $n \to \infty$. Problems are then classified by their *speed of convergence*:

[8] There are only few situations where lower bounds for the Monte Carlo error have been proven directly without switching to the average case setting, see for example the non-adaptive Monte Carlo setting for the integration of univariate monotone functions in Novak [53], or an estimate for small errors for the approximation of monotone Boolean functions in Bshouty and Tamon [9, Thm 5.3.1], see also Remark 4.9.

[9] The notion *complexity setting* comprises all features of algorithms like adaptivity or non-adaptivity, randomization, the class of information functionals $\Lambda$, as well as the error criterion, be it the absolute or the normalized error.

- a function $e(n)$ converges faster than a function $e'(n)$ iff $e(n)/e'(n) \xrightarrow[n\to\infty]{} 0$, we write $e(n) \prec e'(n)$,
- a function $e(n)$ converges at least as fast as a function $e'(n)$ iff there exists a constant $c > 0$ and $n_0 \in \mathbb{N}$ such that $e(n) \le c\, e'(n)$ for $n \ge n_0$, we write $e(n) \preceq e'(n)$,
- two functions $e(n)$ and $e'(n)$ have the same speed of convergence iff $e(n) \preceq e'(n)$ and $e'(n) \preceq e(n)$, we write $e(n) \asymp e'(n)$.[10]

A widely used classification is done by the comparison to polynomial decay, a problem has the *order of convergence* at least $p$ iff $e(n) \preceq n^{-p}$, where $p > 0$. Determining the optimal order of convergence means finding constants $c, C > 0$ such that for large $n$ we have

$$c\, n^{-p} \le e(n) \le C\, n^{-p}.$$

(Sometimes logarithmic factors need to be added.) A common phenomenon when determining the optimal order $p_d$ for $d$-variate problems is that the corresponding constants $c_d$ and $C_d$ deviate widely, and even worse, "large $n$" means $n \ge n_0(d) \in \mathbb{N}$ and $n_0(d)$ can be huge for growing dimension $d$. In Section 4.2.1 we find an example where difficulties become apparent as soon as we consider the inverse notion of $\varepsilon$-complexity. Last but not least, for discrete problems such as the approximation of Boolean functions, see Chapter 4, the concept of order of convergence is meaningless since discrete problems may be solved with a finite amount of information.

For tractability analysis now, we regard the complexity as a function depending on $\varepsilon > 0$ and $d \in \mathbb{N}$,

$$n^{\diamond,\star}(\varepsilon, S^d, F^d, \Lambda) = n(\varepsilon, d)\,.$$

A first approach to this complexity function is to fix $\varepsilon > 0$ and to consider the growth in $d$, see for example the results on lower bounds in Corollary 2.20, Theorem 3.15, or Theorem 4.8. It is unpleasant if the complexity depends exponentially on $d$, we say that a problem suffers from the *curse of dimensionality*[11] iff there exist $\varepsilon, \gamma, c > 0$ and $d_0 \in \mathbb{N}$ such that

$$n(\varepsilon, d) \ge c\,(1+\gamma)^d \quad \text{for } d \ge d_0.$$

There are problems that have arbitrarily high order of convergence but suffer from the curse of dimensionality, see for example the case $p = 1$ in Section 2.4.2.

[10] In Chapter 2 we will encounter relations like $e(n) \ge \frac{1}{2}\, b_{2n}$. It is worth thinking about an alternative definition of *equal speed*, which holds if there exist constants $k_1, k_2, k_3 \in \mathbb{N}$ and $c, C > 0$ such that

$$c\, e(k_1 n) \le e'(k_2 n) \le C\, e(k_3 n)$$

for sufficiently large $n$. For polynomial rates this will not make any difference, but if exponential functions are involved, two functions $\exp(-n^p)$ and $\exp(-2n^p)$ would be classified *the same speed of decay* only for the new notion.

[11] This notion goes back to Bellman 1957 [7].

For positive results, we do not only want the dependency on $d$ to be moderate, but also the dependency on $\varepsilon^{-1}$. A problem is *polynomially tractable* iff there exist constants $C, p, q > 0$ such that

$$n(\varepsilon, d) \leq C\, \varepsilon^{-p}\, d^q\,.$$

If we can even choose $q = 0$, that is if the complexity is essentially independent from the dimension $d$, we have *strong polynomial tractability*.

In contrast to the curse of dimensionality, problems for which the complexity does *not* depend exponentially on $d$ or $\varepsilon^{-1}$, in detail, where

$$\lim_{\varepsilon^{-1}+d\to\infty} \frac{\log n(\varepsilon, d)}{\varepsilon^{-1} + d} = 0\,,$$

are called *weakly tractable*. This notion is fairly new and has been studied first around the time where the book on tractability, Novak and Woźniakowski 2008 [55], has been written. A problem which is *not weakly tractable* is called *intractable*.[12] Note that there are intractable problems that do not suffer from the curse of dimensionality, for example the randomized approximation of monotone functions, see Chapter 4.

More recently, the refined notion of $(s, t)$-*weak tractability* has been promoted in Siedlecki and Weimar [71]. It is fulfilled iff

$$\lim_{\varepsilon^{-1}+d\to\infty} \frac{\log n(\varepsilon, d)}{\varepsilon^{-s} + d^t} = 0$$

with $s, t > 0$. This notion coincides with weak tractability for $s = t = 1$.

Last but not least, Gnewuch and Woźniakowski [21] promoted the notion *quasi-polynomial tractability*. It holds iff there exist constants $C, p > 0$ such that

$$n(\varepsilon, d) \leq C\, \exp[p\,(1 + \log \varepsilon^{-1})(1 + \log d)]\,.$$

In this case the complexity behaves almost polynomially in $d$ with an exponent that grows very slowly in $\varepsilon^{-1}$, and vice versa.

For an example of quasi-polynomial and $(s, t)$-weak tractability, see Theorem 2.21.

Whether or not a problem falls into one of the tractability classes above, highly depends on the particular choice of the $d$-dependent setting $(S^d)_{d\in\mathbb{N}}$. One criterion of a *natural* $d$-dependent problem could be that the input set $F^d$ can be identified with a subset of $F^{d+1}$, and therefore we can consider $S^d$ to be a restriction of $S^{d+1}$.

[12] The notion of "intractability" as it is used within the IBC community since the book on the tractability of multivariate problems, Novak and Woźniakowski [55, p. 14], is different from definitions of "intractability" in other scientific communities. In computer science, see for example the book on *NP-completeness* by Garey and Johnson [19], all problems that, for solving a problem *exactly*, need a running time which is *superpolynomial* in the size $m$ of the input, are called "intractable". Thus even $m^{\log m}$ would fall into that category. In tractability studies for IBC, instead of the input size we consider the dimension $d$ and the error tolerance $\varepsilon$, so automatically new notions arose. But also the observation that many problems have a *sub-exponential* yet *superpolynomial* running time motivated the introduction of new notions like *weak tractability*.

Another possible criterion is whether the initial error is properly *normalized*, that is, the initial error should be a constant,

$$e(0, S^d, F^d) = c > 0 \quad \text{for all } d \in \mathbb{N}.$$

Typically $c = 1$, see for example the problem in Section 2.4.2; however, in Chapter 4 we have $c = \frac{1}{2}$, see Remark 4.6.

## 1.3 Algorithms with Varying Cardinality

For some problems it might be convenient to allow algorithms that collect a varying amount of information, but in average they do not use more than $n$ pieces of information. In Ritter [68, Chap VII and VIII] one can find examples of average case settings where varying cardinality *does* help. Anyway, for upper bounds we try to find algorithms that are as simple as possible, whereas for lower bounds it is desirable that they hold for as general classes of algorithms as possible, that is, we allow for randomization, adaption, or even varying cardinality.[13] In the end one might see what features are really making a big difference.

We need to adjust our model of algorithms $A^\omega = \phi^\omega \circ N^\omega$ where the number of information we collect may depend on the random element $\omega$ and (adaptively) on the input. Now, the information mapping shall be a mapping $N^\omega : F \to \mathbb{R}^{\mathbb{N}}$ yielding an information sequence $\mathbf{y} = (y_k)_{k\in\mathbb{N}}$, and for possible information sequences we need to define an output via a mapping $\phi^\omega : N^\omega(F) \to G$. As before, the $k$-th piece of information is obtained by evaluating an adaptively chosen functional from a given class $\Lambda$, *or the zero functional*,[14]

$$y_k := L^\omega_{k,\mathbf{y}_{[k-1]}}(f)\,.$$

At some point we need to stop collecting further information. Within the model, this means that for some index $n \in \mathbb{N}_0$ we choose $L^\omega_{k,\mathbf{y}_{[k-1]}}$ to be the zero functional for all $k > n$, so the actual amount of information for a particular algorithm is a function

$$n(\omega, \mathbf{y}) := \inf\{n \in \mathbb{N}_0 \mid L^\omega_{k,\mathbf{y}_{[k-1]}} = 0 \text{ for all } k > n\}\,,$$

with $\mathbf{y} := N^\omega(f)$ being a proper information sequence.[15] (For non-adaptive algorithms this function is independent from the input, $n(\omega, \mathbf{y}) = n(\omega)$, for deterministic algorithms it is a function $n(\omega, \mathbf{y}) = n(\mathbf{y})$.) For convenience, we will also write $n(\omega, f)$ instead of $n(\omega, N^\omega(f))$. Then the *worst input cardinality* of the algorithm $A = (A^\omega)_{\omega\in\Omega}$ is defined as

$$\operatorname{card}(A) = \operatorname{card}(A, F) := \sup_{f\in F} \mathbb{E}\, n(\omega, f)\,. \tag{1.3.1}$$

[13] Similarly, it is good to find lower bounds for very small input sets, but upper bounds that hold for very general and large input sets.

[14] Considering for example $\Lambda^{\text{std}}$, in general the zero functional is not a function evaluation.

[15] For fixed $\omega \in \Omega$, not all sequences $\mathbf{y} \in \mathbb{R}^{\mathbb{N}}$ can be the outcome of the information mapping $N^\omega$.

For any (sub)-probability measure $\mu$ the *$\mu$-average cardinality* is

$$\operatorname{card}(A,\mu) := \int \mathbb{E}\, n(\omega, f)\, \mu(\mathrm{d}f)\,. \tag{1.3.2}$$

The $\mu$-average cardinality is usually defined for deterministic algorithms.

As before, we define different classes of algorithms $\mathcal{A}^{\diamond,\star}(\Lambda)$ where $\diamond \in \{\mathrm{ran}, \mathrm{det}\}$ and $\star \in \{\mathrm{ada}, \mathrm{nonada}, \mathrm{lin}, \mathrm{hom}\}$. The definition of the error for a particular algorithm does not change, however, the new concept of cardinality brings about new error and complexity notions associated to a problem $S : F \to G$. For $\bar{n} \geq 0$ we have

$$\bar{e}^{\diamond,\star}(\bar{n}, S, \bullet, \Lambda) := \inf_{\substack{A \in \mathcal{A}^{\diamond,\star}(\Lambda) \\ \operatorname{card}(A,\bullet) \leq \bar{n}}} e(A, S, \bullet)\,,$$

and for a given error tolerance $\varepsilon > 0$ we define

$$\bar{n}^{\diamond,\star}(\varepsilon, S, \bullet, \Lambda) := \inf\{\bar{n} \geq 0 \mid \exists\, A \in \mathcal{A}^{\diamond,\star}(\Lambda) :\ \operatorname{card}(A,\bullet) \leq \bar{n},\ e(A, S, \bullet) \leq \varepsilon\}\,,$$

where for $\bullet$ we may insert an input set $F \subset \widetilde{F}$, or an input distribution $\mu$. Be aware that the cardinality may be a real number now.

Note that algorithms from classes of fixed cardinality $\mathcal{A}_n^{\diamond,\star}(\Lambda)$ can be identified with methods from $\mathcal{A}^{\diamond,\star}(\Lambda)$, so for $\bar{n} \geq 0$ we have the general estimate

$$\bar{e}^{\diamond,\star}(\bar{n}, S, \bullet, \Lambda) \leq e^{\diamond,\star}(\lfloor \bar{n} \rfloor, S, \bullet, \Lambda)\,.$$

For the worst case setting it is easy to see that the new notion even coincides with the old notion of fixed cardinality, that is, for $n \in \mathbb{N}_0$ we have

$$\bar{e}^{\mathrm{det},\star}(n, S, F, \Lambda) = e^{\mathrm{det},\star}(n, S, F, \Lambda)\,.$$

For Monte Carlo methods with non-adaptively varying cardinality $n(\omega)$, there is a direct relation to the fixed cardinality setting. This relation is well known, see Heinrich [22, p. 289/290].

**Lemma 1.4.** *For $n \in \mathbb{N}$ we have*

$$\bar{e}^{\mathrm{ran,nonada}}(n, S, F, \Lambda) \geq \tfrac{1}{2}\, e^{\mathrm{ran,nonada}}(2n, S, F, \Lambda)\,.$$

*Proof.* The proof also works for classes of adaptive algorithms as long as the actual cardinality does not depend on the input. In this sense, let $A = (A^\omega)_{\omega\in\Omega} \in \mathcal{A}^{\mathrm{ran},\star}(\Lambda)$ be a Monte Carlo algorithm with non-adaptively varying cardinality $n(\omega)$ such that $\mathbb{E}\, n(\omega) \leq n \in \mathbb{N}$. Then we have

$$\begin{aligned} e(A, F) &= \sup_{f\in F} \mathbb{E}\, e(A^\omega, S, f) \\ &\geq \sup_{f\in F} \mathbb{E}\, e(A^\omega, S, f)\, \mathbb{1}_{\{n(\omega)\leq 2n\}} \\ &= \mathbb{P}\{n(\omega) \leq 2n\} \sup_{f\in F} \mathbb{E}'\, e(A^\omega, S, f) \end{aligned}$$

Here, $\mathbb{E}'$ denotes the expectation for the conditional probability space $(\Omega', \Sigma \cap \Omega', \mathbb{P}')$ where we integrate over $\omega \in \Omega' := \{\omega \mid n(\omega) \leq 2n\} \subseteq \Omega$ with respect to the conditional measure $\mathbb{P}'(\cdot) := \mathbb{P}(\cdot \mid \Omega') = \mathbb{P}(\cdot \cap \Omega')/\mathbb{P}(\Omega')$. We can regard $A' := (A^\omega)_{\omega\in\Omega'}$ as a Monte Carlo algorithm from the class $\mathcal{A}_{2n}^{\mathrm{ran},\star}(\Lambda)$ with another underlying probability space than for $A$. Together with $\mathbb{P}\{n(\omega) \leq 2n\} \geq \frac{1}{2}$ (by Markov's inequality), this gives the lower bound

$$e(A, F) \geq \tfrac{1}{2}\, e^{\mathrm{ran},\star}(2n, S, F, \Lambda)\,.$$

□

For Monte Carlo methods with adaptively varying cardinality $n(\omega, f)$ we need special versions of Bakhvalov's technique.

**Lemma 1.5** (Bakhvalov's technique for varying cardinality)**.** *For any (sub-)probability measure $\mu$ on $F$, and $\bar{n} \geq 0$, the Monte Carlo error and the average error in the setting of (adaptively) varying cardinality are related by*

$$\bar{e}^{\mathrm{ran,ada}}(\bar{n}, S, F, \Lambda) \geq \tfrac{1}{2}\, \bar{e}^{\mathrm{avg,ada}}(2\bar{n}, S, \mu, \Lambda)\,.$$

*If we have an estimate $\bar{e}^{\mathrm{avg,ada}}(\bar{n}, S, \mu, \Lambda) \geq \hat{\varepsilon}(\bar{n})$ with a convex and decaying function $\hat{\varepsilon}(\bar{n})$ for $\bar{n} \geq 0$, the lower bound can be improved to*

$$\bar{e}^{\mathrm{ran,ada}}(\bar{n}, S, F, \Lambda) \geq \hat{\varepsilon}(\bar{n})\,.$$

*Proof.* Let $A = (A^\omega)_{\omega\in\Omega} \in \mathcal{A}^{\mathrm{ran},\star}(\Lambda)$ be a Monte Carlo algorithm with adaptively varying cardinality $n(\omega, f)$ such that

$$\bar{n} \geq \mathrm{card}(A, F) := \sup_{f\in F} \mathbb{E}\, n(\omega, f)\,.$$

We can relate this to the average cardinality with respect to $\mu$ regarding $A^\omega$ as a deterministic algorithm for fixed $\omega$,

$$\begin{aligned}
&\geq \int \mathbb{E}\, n(\omega, f)\, \mu(\mathrm{d}f) \\
\text{[Fubini]} \quad &= \mathbb{E} \underbrace{\int n(\omega, f)\, \mu(\mathrm{d}f)}_{=\mathrm{card}(A^\omega, \mu)} \\
\text{[Markov's ineq.]} \quad &\geq 2\bar{n}\, \mathbb{P}\{\mathrm{card}(A^\omega, \mu) > 2\bar{n}\}\,.
\end{aligned}$$

This gives us the estimate

$$\mathbb{P}\{\mathrm{card}(A^\omega, \mu) \leq 2\bar{n}\} \geq \tfrac{1}{2}\,. \tag{1.3.3}$$

Now, considering the error, we find

$$\begin{aligned}
e(A, F) &= \sup_{f\in F} \mathbb{E}\, e(A^\omega, f) \\
&\geq \int \mathbb{E}\, e(A^\omega, f)\, \mu(\mathrm{d}f) \\
\text{[Fubini]} \quad &= \mathbb{E} \int e(A^\omega, f)\, \mu(\mathrm{d}f) \\
&= \mathbb{E}\, e(A^\omega, \mu) \\
&\geq \mathbb{E}\, \bar{e}^{\mathrm{avg},\star}(\mathrm{card}(A^\omega, \mu), S, \mu, \Lambda)\,.
\end{aligned}$$

A rough estimate via (1.3.3) will give

$$e(A,F) \geq \underbrace{\mathbb{P}\{\operatorname{card}(A^\omega,\mu) \leq 2\bar{n}\}}_{=\frac{1}{2}}\, \bar{e}^{\,\mathrm{avg},\star}(2\bar{n},S,\mu,\Lambda)\,.$$

If we have a specially structured estimate $\hat{\varepsilon}$ for the average error, we can proceed in a better way,

$$\begin{aligned} e(A,F) &\geq \mathbb{E}\,\hat{\varepsilon}(\operatorname{card}(A^\omega,\mu)) \\ \text{[convexity]} \quad &\geq \hat{\varepsilon}(\mathbb{E}\operatorname{card}(A^\omega,\mu)) \\ \text{[monotonicity]} \quad &\geq \hat{\varepsilon}(\bar{n})\,. \end{aligned}$$

This finishes the proof. □

A similar convexity argument will help to find good bounds for average case settings with varying cardinality.

**Lemma 1.6** (Special average settings with varying cardinality)**.** *Let $\mu$ be a probability measure on $\widetilde{F}$. Assume that for any deterministic algorithm $\phi \circ N$ with varying cardinality there exists a version of the conditional measure $\mu_{\mathbf{y}}$ such that*

$$\int \operatorname{dist}_G(\phi(\mathbf{y}),S(f))\,\mu_{\mathbf{y}}(\mathrm{d}f) \geq \hat{\varepsilon}(n(\mathbf{y}))\,,$$

*where $\hat{\varepsilon}(\bar{n})$ is convex and decaying for $\bar{n} \geq 0$. Then the average error for algorithms with varying cardinality is bounded by this function,*

$$\bar{e}^{\,\mathrm{avg,ada}}(\bar{n},S,\mu,\Lambda) \geq \hat{\varepsilon}(\bar{n})\,.$$

*If $\mu$ is supported on $F$, by Lemma 1.5 the very lower bound holds for the Monte Carlo error as well.*

*Proof.* Let $A = \phi \circ N$ be a deterministic algorithm with adaptively varying cardinality $n(f)$ such that $\bar{n} \geq \int n(f)\,\mu(\mathrm{d}f)$. By definition we have

$$e(A,\mu) = \int e(A,f)\,\mu(\mathrm{d}f)\,.$$

We split the integral into the integration over $\mathbf{y} \in N^\omega(F)$, with an appropriate conditional distributions $\mu_{\mathbf{y}}$ on $N^{-1}(\mathbf{y}) \cap F$, fulfilling the assumptions of the lemma, and obtain

$$\begin{aligned} &= \int \left[\int \operatorname{dist}_G(\phi(\mathbf{y}),S(f))\,\mu_{\mathbf{y}}(\mathrm{d}f)\right] \mu \circ N^{-1}(\mathrm{d}\mathbf{y}) \\ &\geq \int \hat{\varepsilon}(n(\mathbf{y}))\,\mu \circ N^{-1}(\mathrm{d}\mathbf{y}) \\ \text{[convexity]} \quad &\geq \hat{\varepsilon}\left(\int n(f)\,\mu(\mathrm{d}f)\right) \\ \text{[monotonicity]} \quad &\geq \hat{\varepsilon}(\bar{n})\,. \end{aligned}$$

□

By Lemma 1.4 we see that non-adaptively varying cardinality does not help a lot when trying to find better Monte Carlo algorithms. For adaptively varying cardinality the situation is slightly more complicated; however, Lemma 1.5 gives us a tool to prove lower bounds that are similar to those that we can obtain for the fixed cardinality setting, see Section 2.3.1 for an application of this lemma. In many more cases even better, Lemma 1.6 applies to the average setting so that we obtain lower bounds which coincide with the computed estimates for the fixed cardinality setting. In this dissertation, we have this nice situation for the lower bounds in Theorem 2.16 (homogeneous algorithms and Bernstein numbers), and in Theorem 4.2 and Theorem 4.8 (approximation of monotone functions). This justifies that in the main parts of this thesis we focus on algorithms with fixed cardinality.

## 1.4 On the Measurability of Algorithms

It seems *natural* to assume measurability for algorithms since *real computers* can only deal with a finite amount of states. In the IBC setting, however, it is convenient to assume that we can operate with *real numbers*, otherwise the concept of linear algorithms for real-valued functions would not make sense. Further justification for why we work with the real number model is gathered in Novak and Woźniakowski [55, Sec 4.1.3]. Unfortunately, the *real number model* tails the problem of measurability. Heinrich and Milla [26] presented a simple Monte Carlo sampling algorithm for indefinite integration that at first view appears natural but, in fact, is not measurable.[16] We will comment on that.

Consider the indefinite integration

$$S^d : L_p([0,1]^d) \to L_\infty([0,1]^d), \quad [S^d(f)](\mathbf{x}) := \int_{[\![\mathbf{0},\mathbf{x}]\!]} f \,\mathrm{d}\lambda^d \,,$$

where $1 < p \leq \infty$. The simple Monte Carlo sampling algorithm $A_n$ is given by

$$[A_n(f)](\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\mathbf{X}_i \leq \mathbf{x}] \, f(\mathbf{X}_i) \,,$$

with iid random variables $\mathbf{X}_i \sim \mathrm{unif}([0,1]^d)$. As discussed in [26, Sec 6.3], this algorithm is not measurable since the method is not separably valued. Indeed, considering the constant function $f_1 = 1$, for two realizations $A_n^\omega$ and $A_n^{\omega'}$ with distinct sample points $\mathbf{X}_i(\omega)$ and $\mathbf{X}_i(\omega')$ modulo ordering, we have $\|A_n^\omega f_1 - A_n^{\omega'} f_1\|_\infty \geq 1/n$. Still, the error mapping $\omega \mapsto \|S^d(f) - A_n^\omega(f)\|_\infty$ is measurable and an error analysis makes sense.[17] In detail, Heinrich and Milla show that it suffices to consider the pointwise

[16] I would like to thank Mario Hefter for interesting discussions on measurability of algorithms during our stay at Brown University's ICERM in fall 2014. I would also like to thank Prof. Dr. Klaus Ritter for pointing me to the paper of Heinrich and Milla [26].

[17] In detail, Heinrich and Milla [26, Thm 3.4] showed polynomial tractability in the randomized setting. Note that in the deterministic setting the problem is unsolvable because we may only use function values of $L_p$-functions as information. By this, indefinite integration is an example

difference $|[S^d(f)](\mathbf{x}) - [A_n^\omega(f)](\mathbf{x})|$ for points $\mathbf{x}$ from a regular grid $\Gamma_m \subset [0,1]^d$ with mesh size $1/m$, see [26, Sec 3].

This motivates a very natural measurable modification of the sample algorithm. A computer can only store finitely many digits of the coordinates of $\mathbf{X}_i$, in general, for fixed $\omega$, the function $g_i^\omega(\mathbf{x}) := \mathbb{1}[\mathbf{X}_i \le \mathbf{x}]$ is not exactly implementable. Therefore, let $\mathbf{X}_i^{(r)}(j) \le \mathbf{X}_i(j)$ be the largest rational number representable with $r$ binary digits after the radix point, $j = 1, \dots, d$. The algorithm

$$[A_{n,r}(f)](\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\mathbf{X}_i^{(r)} \le \mathbf{x}]\, f(\mathbf{X}_i)$$

is composed of measurable mappings, and thereby measurable itself. Indeed, the mapping

$$\Omega \to L_\infty([0,1]^d), \quad \omega \mapsto \mathbb{1}[(\mathbf{X}_i^{(r)})^\omega \le \cdot\,]$$

is measurable since it only has discrete values. Furthermore, the pointwise error of $A_n$ and $A_{n,r}$ coincides on the grid $\Gamma_{2^r}$. With increasing $r$ we can get arbitrarily close to the error of $A_n$, compare Heinrich and Milla [26, Thm 3.4]. The original publication contains another modification with continuous outputs. The modification given here, however, nourishes the belief in measurability of implementable algorithms.

This was an example of non-measurability of Monte Carlo algorithms. Typically, measurability is an assumption in the randomized and in the average setting, but we do not need it for the worst case setting. If we assume measurability only for randomized algorithms, but allow non-measurability for deterministic algorithms in the worst case setting, for linear problems $S$ with the input set $F$ being the unit ball in $\widetilde{F}$ and with general linear information $\Lambda^{\mathrm{all}}$, one can still state

$$\bar{e}^{\mathrm{ran}}(n, S, F, \Lambda^{\mathrm{all}}) \le 4\, e^{\mathrm{det}}(n, S, F, \Lambda^{\mathrm{all}})\,,$$

see Heinrich [22, p. 282, (4)].

As we have seen in the example above, we do not really need measurability for the algorithm as long as the error mapping is measurable. Measurability, however, is a convenient assumption, especially for the average case analysis in Chapter 2, where we need to establish the conditional measure for given information $\mathbf{y}$, see Section 2.2.1. As long as we do not find meaningful non-measurable algorithms that could not be replaced by equally successful measurable algorithms, measurability is a justifiable assumption for lower bound studies. For the lower Monte Carlo bounds in Chapter 4, however, measurability is unproblematic since we consider average settings with discrete measures, see Theorem 4.2 and Theorem 4.8. In that case, relaxed measurability assumptions would suffice, but we do not go into details.

---

of a problem where the output space consists of functions and where randomization does help. Heinrich and Milla also note that only few polynomially tractable problems with unweighted dimensions have been known so far. Their example of indefinite integration is such an unweighted problem with polynomial tractability. In Section 3.4.2 of this dissertation we add another example: The $L_\infty$-approximation of Hilbert space functions from unweighted periodic Korobov spaces with standard information $\Lambda^{\mathrm{all}}$ is polynomially tractable in the Monte Carlo setting.

We finish with a final remark on an alternative approximation concept, aside from the IBC setting with the real number model. Given a numerical problem $S : F \to G$, we define *entropy numbers* for $n \in \mathbb{N}_0$,

$$e_n(S,F) := \inf\left\{\varepsilon > 0 \,\middle|\, \exists g_1, \ldots, g_{2^n} \in G : S(F) \subseteq \bigcup_{i=1}^{2^n} B_G(g_i, \varepsilon)\right\},$$

where $B_G(g,\varepsilon)$ denotes the closed $\varepsilon$-ball around $g \in G$, see Carl [10], alternatively Pisier [66, Chap 5]. One interpretation of this concept is the question on how well we can approximate the problem $S$ if we are only allowed to use $n$ bits to represent $2^n$ different outputs.[18] Carl studies linear problems and establishes a lower bound for certain *s-numbers* based on entropy numbers. Some of the $s$-numbers are closely related to the error quantities for deterministic algorithms with $\Lambda^{\mathrm{all}}$, *approximation numbers* correspond to the error of linear methods, *Gelfand numbers* are linked to the error of general deterministic methods.[19]
Within this dissertation, in Remark 4.9 we cite a lower bound for the Monte Carlo error for the approximation of monotone functions which is due to Bshouty and Tamon [9]. Their proof uses an entropy argument.
In Chapter 3, in the context of estimates on the expected maximum of zero-mean Gaussian fields, we will step across the inverse concept *metric entropy* $H(\varepsilon)$, that is the logarithm of the minimal number of $\varepsilon$-balls needed to cover a set, see Proposition 3.7 (Dudley).

[18] The given definition contains an index shift compared to the definition to be found in Carl [10]. This is a matter of taste. Here, $e_0(S,F)$ coincides with the initial error from the IBC setting. According to Carl's notation, we would start with $n = 1$, and the initial error would match $e_1(S,F)$. Similar index shifts compared to related notions from IBC are commonly found for *s-numbers*, following the axiomatic scheme of Pietsch [63]. Contrarily, Hutton, Morrell, and Retherford [29] use a definition of approximation numbers which happens to fit the IBC notion. Heinrich [22], in turn, in his paper on lower bounds, on which Chapter 2 is based on, and Mathé [46] in his fundamental research on random approximation by $\Lambda^{\mathrm{all}}$, which inspired Chapter 3, both kept consistency with the $s$-number conventions, even for the definition of the Monte Carlo error. In this thesis, however, we strictly follow IBC conventions for error quantities. In contrast, for the definition of Bernstein numbers in Section 2.1, we use a definition which fits to the $s$-number scheme, see the footnote given there for additional justification. See also Lemma 3.9 (b) for the link between singular values and the worst case error in the Hilbert space setting.

[19] See the book on IBC by Traub et al. [73, pp. 70–73].

# Chapter 2

# Lower Bounds for Linear Problems via Bernstein Numbers

We consider adaptive Monte Carlo methods for linear problems and establish a lower bound via Bernstein numbers, see Section 2.1 for the definition and an overview of already known relations. The abstract main result and the proof is contained in Section 2.2. It is based on a technique due to Heinrich [22] that relates the Monte Carlo error to norm expectations of Gaussian measures. The innovation is the application of Lewis' theorem in order to find optimal Gaussian measures, see Section 2.2.3. Within the supplementary Section 2.3 we present versions of the main result for two interesting special settings: varying cardinality, and homogeneous algorithms. A major application is the $L_\infty$-approximation of certain classes of $C^\infty$-functions, see Section 2.4.2. With the new technique we obtain lower bounds via Bernstein numbers, which show that in these cases randomization cannot give us better tractability than that what we already have with deterministic methods.

## 2.1 The Setting and Bernstein Numbers

Let $S : \widetilde{F} \to G$ be a compact linear operator between Banach spaces over the reals. Throughout this chapter the *input set* $F \subset \widetilde{F}$ is the unit ball of $\widetilde{F}$, the corresponding norm is denoted by $\|\cdot\|_F$. We consider algorithms that may use arbitrary continuous linear functionals $\Lambda^{\mathrm{all}}$ as information.

The operator $S$ can be analysed in terms of *Bernstein numbers*

$$b_m(S) := \sup_{X_m \subseteq \widetilde{F}} \inf_{\substack{f \in X_m \\ \|f\|=1}} \|S(f)\|_G \,, \tag{2.1.1}$$

where the supremum is taken over $m$-dimensional[1] linear subspaces $X_m \subseteq \widetilde{F}$. These

[1]Some authors take the supremum over $(m+1)$-dimensional spaces [49, 61], which might be motivated by relations like (2.1.3). The present version, however, is also in common use [38, 50], besides it looks quite natural, and in view of the sharp estimate (2.1.7), any index shift would appear like a disimprovement. Although Bernstein numbers are not $s$-numbers, according to the definition in Pietsch [63], in some cases they coincide with certain $s$-numbers, in particular for operators between Hilbert spaces where Bernstein numbers match the singular values.

quantities are closely related to the *Bernstein widths*[2] of the image $S(F)$ within $G$,

$$b_m(S(F),G) := \sup_{Y_m \subseteq G} \sup\{r \geq 0 \mid B_r(0) \cap Y_m \subseteq S(F)\}\,, \tag{2.1.2}$$

where the first supremum is taken over $m$-dimensional linear subspaces $Y_m \subseteq G$. By $B_r(g)$ we denote the (closed) ball around $g \in G$ with radius $r$. In general, Bernstein widths are greater than Bernstein numbers, however, for injective operators (like embeddings) both notions coincide (consider $Y_m = S(X_m)$). In the case of Hilbert spaces $\widetilde{F}$ and $G$, Bernstein numbers and widths match the singular values $\sigma_m(S)$.

For deterministic algorithms it can be easily seen that

$$e^{\mathrm{det}}(n,S) \geq b_{n+1}(S(F),G) \geq b_{n+1}(S)\,, \tag{2.1.3}$$

since for any information mapping $N : \widetilde{F} \to \mathbb{R}^n$ and any $\varepsilon > 0$, there always exists an $f \in N^{-1}(\mathbf{0})$ with $\|S(f)\|_G \geq b_{n+1}(S(F),G) - \varepsilon$ and $\pm f \in F$, i.e. $f$ cannot be distinguished from $-f$.

If both $\widetilde{F}$ and $G$ are Hilbert spaces, lower bounds for the (root mean square) Monte Carlo error have been found by Novak [52],

$$e_2^{\mathrm{ran}}(n,S) \geq \tfrac{\sqrt{2}}{2}\,\sigma_{2n}(S)\,. \tag{2.1.4}$$

For operators between arbitrary Banach spaces the estimate reads quite similar, see Theorem 2.1,

$$e^{\mathrm{ran,ada}}(n,S) > \tfrac{1}{30}\,b_{2n}(S)\,. \tag{2.1.5}$$

The constant can be improved for extremely large $n$, see Remark 2.13, or when imposing further assumptions. The following lower bound for non-adaptive algorithms has been proven first within the author's master thesis and published later in [39],

$$e^{\mathrm{ran,nonada}}(n,S) \geq \tfrac{1}{2}\,b_{2n+1}(S)\,. \tag{2.1.6}$$

For homogeneous algorithms, possibly adaptive, and even with varying cardinality, one can prove an estimate with optimal constant, see Theorem 2.16,

$$e^{\mathrm{ran,hom}}(n,S) \geq \tfrac{1}{2}\,b_{2n}(S)\,. \tag{2.1.7}$$

Within [39] the results of (2.1.5) and (2.1.7) have been mentioned, for the adaptive setting a proof for a result with slightly worse constants based on results from Heinrich [22] has been given. In this chapter now one can find a self-contained proof following the lines of Heinrich [22] but with optimized constants and slight simplifications that are possible when relying on Bernstein numbers.

[2] I wish to thank Prof. Dr. Stefan Heinrich for making me aware of the non-equivalence of both notions.

## 2.2 Adaptive Monte Carlo Methods

Theorem 2.1 below is the main result of this chapter. The proof needs several results that are provided in the subsequent subsections.

The proof is based on the idea of Heinrich [22] to use truncated Gaussian measures in order to obtain lower bounds for the Monte Carlo error. Considering Gaussian measures is quite convenient as there is an easy representation for the conditional distribution, even when collecting adaptive information, see Section 2.2.1. The key tool for Heinrich's technique is a deviation result for zero-mean Gaussian measures $\tilde{\mu}$ on a normed space $\widetilde{F}$,

$$\tilde{\mu}\{f \,:\, \|f\|_F > \lambda \, \mathbb{E}^{\tilde{\mu}} \|f\|_F\} \leq \exp\left(-\frac{(\lambda-1)^2}{\pi}\right),$$

see Corollary A.6. This shows how far the norm may deviate from its expected value, that way enabling us to estimate how much we lose when truncating a Gaussian measure. The expected norm for a truncated Gaussian measure is estimated in Section 2.2.2, in the case of Bernstein numbers a simplified result with slightly better constants is feasible.

The new idea now is to apply Lewis' theorem in order to find optimal Gaussian measures that are "well spread" into all directions within the input space $\widetilde{F}$, see Section 2.2.3.[3]

This dissertation includes a self-contained proof of the theorem. That way we are able to adapt for simplifications that are possible in the particular situation. Furthermore, we work on the improvement of constants.

**Theorem 2.1.** *For $S : \widetilde{F} \to G$ being a compact linear operator between Banach spaces, and the input set $F$ being the unit ball in $\widetilde{F}$, we have*

$$e^{\mathrm{ran,ada}}(n, S, F, \Lambda^{\mathrm{all}}) > \frac{1}{15} \frac{m-n}{m} \, b_m(S) \quad \textit{for } m > n.$$

*Proof.* For all $\varepsilon > 0$ there exists an $m$-dimensional subspace $X_m \subseteq \widetilde{F}$ such that

$$\|S(f)\|_G \geq \|f\|_F \, (b_m(S) - \varepsilon) \quad \text{for } f \in X_m.$$

Note that for the restricted operator we have $b_m(S|_{X_m}) \geq b_m(S) - \varepsilon$, and in general $e^{\mathrm{ran}}(n, S, F) \geq e^{\mathrm{ran}}(n, S|_{X_m}, F \cap X_m)$. Hence it suffices to show the theorem for $S|_{X_m}$, so without loss of generality we assume $X_m = \widetilde{F} = \mathbb{R}^m$, and therefore $\|S(f)\|_G \geq \|f\|_F \, b_m(S)$ holds for all $f \in \widetilde{F}$.

Below, $\mathbb{P}$ and $\mathbb{E}$ are used to describe probabilities and expectations for an average case setting whenever it seems convenient. This is not to be confused with the probability space $(\Omega, \Sigma, \mathbb{P})$ used to define Monte Carlo algorithms within Chapter 1. Let $\mathbf{X}$ be a standard Gaussian vector within $\mathbb{R}^m = \ell_2^m$. We choose a

[3] This idea has already been published in [39]. I wish to thank Prof. Dr. Aicke Hinrichs and my doctoral advisor Prof. Dr. Erich Novak for pointing me to the book of Pisier [66] in search of optimal Gaussian measures.

matrix $J : \mathbb{R}^m = \ell_2^m \to \mathbb{R}^m = \widetilde{F}$ in order to define a Gaussian measure $\tilde{\mu}$ on $\widetilde{F}$ as the distribution of $J\mathbf{X}$. The restricted measure

$$\mu(E) := \tilde{\mu}|_F(E) = \mathbb{P}\{J\mathbf{X} \in E \cap F\}\,, \quad \text{for measurable } E \subseteq \widetilde{F},$$

is a sub-probability measure supported on the unit ball $F \subset \widetilde{F}$. By Bakhvalov's technique, see Proposition 1.2, we know

$$e^{\mathrm{ran,ada}}(n, S, F, \Lambda^{\mathrm{all}}) \geq e^{\mathrm{avg,ada}}(n, S, \mu, \Lambda^{\mathrm{all}})\,.$$

Let $\phi \circ N : \widetilde{F} \to G$ be an adaptive deterministic algorithm using $n$ pieces of information. Let $\tilde{\nu} = \tilde{\mu} \circ N^{-1}$ denote the distribution of the information $N(J\mathbf{X})$. Without loss of generality, $N : \widetilde{F} \to \mathbb{R}^n$ is surjective, so by Lemma 2.2, for all $\mathbf{y} \in \mathbb{R}^n$ we have an orthogonal projection $P_{\mathbf{y}}$ and an element $m_{\mathbf{y}} \in \widetilde{F}$ to describe the conditional distribution $\tilde{\mu}_{\mathbf{y}}$ as the distribution of $JP_{\mathbf{y}}\mathbf{X} + m_{\mathbf{y}}$. We write the error

$$e(\phi \circ N, \mu) \;= \int_{\mathbb{R}^n} \int_{F \cap N^{-1}(\mathbf{y})} \|S(f) - \phi(\mathbf{y})\|_G \, \tilde{\mu}_{\mathbf{y}}(\mathrm{d}f)\, \tilde{\nu}(\mathrm{d}\mathbf{y})\,.$$

Defining $g_{\mathbf{y}} := \phi(\mathbf{y}) - S(m_{\mathbf{y}})$, we can continue using the representation of the conditional measure $\tilde{\mu}_{\mathbf{y}}$, further cutting off parts of the integral,

$$\geq \int_{\mathbb{R}^n} \mathbb{E}\Big[ \|SJP_{\mathbf{y}}\mathbf{X} - g_{\mathbf{y}}\|_G \, \mathbb{1}_{\{\|JP_{\mathbf{y}}\mathbf{X}\|_F \leq 1 - \|m_{\mathbf{y}}\|_F\}} \Big]\, \tilde{\nu}(\mathrm{d}\mathbf{y})\,.$$

Due to symmetry, the two versions $\|SJP_{\mathbf{y}}\mathbf{X} \pm g_{\mathbf{y}}\|_G \, \mathbb{1}_{\{\|JP_{\mathbf{y}}\mathbf{X}\|_F \leq r\}}$ are identically distributed ($r > 0$), so we can rewrite

$$= \int_{\mathbb{R}^n} \mathbb{E}\Big[ \underbrace{\frac{1}{2} \sum_{\sigma = \pm 1} \|SJP_{\mathbf{y}}\mathbf{X} + \sigma g_{\mathbf{y}}\|_G}_{\substack{[\Delta\text{-ineq.}] \quad \geq \|SJP_{\mathbf{y}}\mathbf{X}\|_G}} \, \mathbb{1}_{\{\|JP_{\mathbf{y}}\mathbf{X}\|_F \leq 1 - \|m_{\mathbf{y}}\|_F\}} \Big]\, \tilde{\nu}(\mathrm{d}\mathbf{y})\,.$$

Applying the triangle inequality, and further truncating the integral, we get

$$\geq \int_{\{\|m_{\mathbf{y}}\|_F \leq 1 - r\}} \mathbb{E}\Big[ \|SJP_{\mathbf{y}}\mathbf{X}\|_G \, \mathbb{1}_{\{\|JP_{\mathbf{y}}\mathbf{X}\|_F \leq r\}} \Big]\, \tilde{\nu}(\mathrm{d}\mathbf{y})\,.$$

Using the definition of the Bernstein numbers, and replacing the projections $P_{\mathbf{y}}$ by a general estimate for orthogonal rank-$(m-n)$ projections, we end up with

$$e(\phi \circ N, \mu) \;\geq\; \tilde{\nu}\{\mathbf{y} \,:\, \|m_{\mathbf{y}}\|_F \leq 1 - r\} \inf_{\substack{P \text{ orth. Proj.} \\ \mathrm{rk}\, P = m - n}} \mathbb{E}\Big[ \|JP\mathbf{X}\|_F \, \mathbb{1}_{\{\|JP_{\mathbf{y}}\mathbf{X}\|_F \leq r\}} \Big]\, b_m(S)\,.$$

From now on we write $\alpha := \mathbb{E}\,\|J\mathbf{X}\|_F$.

First, we need an estimate for the probability of a small $m_{\mathbf{y}}$, this is done in Lemma 2.3 with $1 - r = 2\,\kappa\,\alpha$,

$$\tilde{\nu}\{\mathbf{y} \,:\, \|m_{\mathbf{y}}\|_F \leq 1 - r\} \geq 1 - 2\,\exp\left( - \left( \frac{1 - r}{2\,\alpha} - 1 \right)^2 \Big/ \pi \right) =: \nu(r, \alpha)\,. \qquad (2.2.1)$$

This estimate is meaningful for $r < 1 - 2\,\alpha\,(1 + \sqrt{\pi \log 2})$.

Second, for orthogonal projections $P$ on $\ell_2^m$ and $t > 0$, the truncated expectation can be estimated by

$$\begin{aligned} \mathbb{E}\Big[\|JP\mathbf{X}\|_F \mathbb{1}_{\{\|JP\mathbf{X}\|_F \le \lambda\, \mathbb{E}\,\|J\mathbf{X}\|_F\}}\Big] &\ge \mathbb{E}\Big[\|JP\mathbf{X}\|_F \mathbb{1}_{\{\|JP\mathbf{X}\|_F \le \lambda\, \mathbb{E}\,\|JP\mathbf{X}\|_F\}}\Big] \\ &\ge \beta(\lambda)\, \mathbb{E}\,\|JP\mathbf{X}\|_F\,. \end{aligned} \tag{2.2.2}$$

Here, within the first step we used $\mathbb{E}\,\|JP\mathbf{X}\|_F \le \mathbb{E}\,\|J\mathbf{X}\|_F$, see Lemma A.3. The second inequality is the application of Lemma 2.5 with the operator $J' := JP$ and the constant $\beta(\lambda)$ defined there. In our situation $\lambda = \frac{r}{\alpha}$.

By Corollary 2.9 we know that $J$ can be chosen in a way such that for any rank $(m-n)$ projection $P$ on $\mathbb{R}^m$ we have[4]

$$\mathbb{E}\,\|JP\mathbf{X}\|_F \ge \frac{m-n}{m}\, \mathbb{E}\,\|J\mathbf{X}\|_F\,.$$

In detail, we choose $J := \alpha\widetilde{J}$, with $\widetilde{J}$ being the optimal operator from Corollary 2.9, and $\alpha > 0$.

Putting all this together, we obtain the estimate

$$e(\phi \circ N, \mu) \ge c\,\frac{m-n}{m}\, b_m(S)$$

with $c := \nu(r,\alpha)\,\beta\left(\frac{r}{\alpha}\right)\,\alpha$. Note that $\nu(r,\alpha) > 0$ iff $r < 1 - 2\,\alpha\,(1 + \sqrt{\pi \log 2})$. On the other hand, $\beta(\lambda)$ gives meaningful results for $\lambda > 3.0513$ only, so we need $r > 3.0513\,\alpha$ in order to obtain positive estimates. Combining this, we have the constraint

$$0 < \alpha < \frac{1}{5.0513 + 2\sqrt{\pi \log 2}} < 0.125 = \frac{1}{8}\,.$$

With $r = 0.37$ and $\alpha = 0.0735$, we find a constant $c = 0.06667... > \frac{1}{15}$. □

### 2.2.1 The Conditional Measure

The following lemma gives the conditional measure for adaptive information mappings applied in an Gaussian average case setting. The conditional measure is well known since the study of average errors, see the book on IBC, Traub et al. [73, pp. 471]. This reference has also been given in Heinrich [22, p. 287]. The proof given here is intended to be self-contained and uses a slightly different notation.

**Lemma 2.2.** *Let* $\mathbf{X}$ *be a standard Gaussian vector in* $\mathbb{R}^m$*, and* $J : \mathbb{R}^m \to \widetilde{F}$ *an injective linear operator defining a measure* $\tilde{\mu}$ *on* $\widetilde{F}$ *as the distribution of* $f := J\mathbf{X}$*. Furthermore, let* $N : \widetilde{F} \to \mathbb{R}^n$ *be a* non-wasteful[5] *adaptive deterministic information*

[4] This idea is new and special for the situation of Bernstein numbers. However, similar properties have been known to Heinrich [22, Lem 3] for the special case of the standard Gaussian distribution in sequence spaces $\ell_p^m$, compare also Remark 2.10. Heinrich used a symmetry argument that can be found in Mathé [46, Lem 4].

[5] That means, $N(\mathrm{supp}(\tilde{\mu})) = \mathbb{R}^n$, so if $\widetilde{F}$ is $m$-dimensional, $N$ shall be surjective.

*mapping. Then the conditional measure $\tilde{\mu}_{\mathbf{y}}$, given the information $\mathbf{y} = N(f)$, can be described as the distribution of $JP_{\mathbf{y}}\mathbf{X} + m_{\mathbf{y}}$, with $P_{\mathbf{y}}$ being a suitable rank-$(m-n)$ orthogonal projection within $\mathbb{R}^m = \ell_2^m$, and a suitable vector $m_{\mathbf{y}} \in \widetilde{F}$. That is, for all measurable $E \subseteq \widetilde{F}$ we have*

$$\tilde{\mu}(E) = \mathbb{P}\{J\mathbf{X} \in E\} = \int_{\mathbb{R}^n} \underbrace{\mathbb{P}\{JP_{\mathbf{y}}\mathbf{X} + m_{\mathbf{y}} \in E\}}_{=\tilde{\mu}_{\mathbf{y}}(E)} \, \tilde{\mu} \circ N^{-1}(\mathrm{d}\mathbf{y}) \,.$$

*Proof.* Before going into the details of the proof, we want to clarify that expressions containing $\mathbf{X}$ are random variables, whereas the information vector $\mathbf{y}$, and everything depending on it, is fixed. In particular, $N(J\mathbf{X})$ is the information as a random vector, $\{N(J\mathbf{X}) = \mathbf{y}\}$ is an event fixing the information.

We denote the partial information $N_k(f) := \mathbf{y}_{[k]}$ for $k = 0, 1, \dots, n$. We will show by induction that the conditional measure $\tilde{\mu}_{\mathbf{y}_{[k]}}$, knowing the first $k$ information values $\mathbf{y}_{[k]}$, can be represented as the distribution of $JP_{k,\mathbf{y}_{[k-1]}}\mathbf{X} + m_{\mathbf{y}_{[k]}}$, with $P_{k,\mathbf{y}_{[k-1]}}$ being a suitable rank-$(m-k)$ orthogonal projection within $\mathbb{R}^m = \ell_2^m$, and a suitable vector $m_{\mathbf{y}_{[k]}} \in \widetilde{F}$. Moreover, there exists a vector $\mathbf{n}_{\mathbf{y}_{[k]}} \in \ell_2^m$ such that $m_{\mathbf{y}_{[k]}} = J\mathbf{n}_{\mathbf{y}_{[k]}}$ and $P_{k,\mathbf{y}_{[k-1]}}\mathbf{n}_{\mathbf{y}_{[k]}} = \mathbf{0}$. For convenience, we also show that the information mapping can be chosen in a way such that the distribution $\tilde{\mu} \circ N_k^{-1}$ of the partial information $N_k(J\mathbf{X}) = \mathbf{y}_{[k]}$ is the $k$-dimensional standard Gaussian distribution which we denote by $\gamma_k$.

Starting with $k = 0$ means that we have no information $\mathbf{y}_\emptyset = 0 \in \mathbb{R}^0$,[6] the conditional distribution $\tilde{\mu} = \tilde{\mu}_{\mathbf{y}_\emptyset}$ is described by $J\mathbf{X} = JP_{0,\mathbf{y}_\emptyset}\mathbf{X} + m_0$ where $P_{0,\mathbf{y}_\emptyset} = \mathrm{id}_{\ell_2^m}$ and $m_0 = 0 \in \widetilde{F}$, or $\mathbf{n}_0 = \mathbf{0} \in \ell_2^m$. The partial information is distributed according to the "zero-dimensional standard Gaussian distribution", that is $\mathbb{P}\{N_0(J\mathbf{X}) = 0\} = 1$.

Now for $k = 1, \dots, n$. Given the partial information $\mathbf{y}_{[k-1]}$, the (adaptively chosen) $k$-th information functional $L_{k,\mathbf{y}_{[k-1]}}$ actually gives us information about the random vector $\mathbf{X} \in \mathbb{R}^m$. In detail, $L_{k,\mathbf{y}_{[k-1]}}(J\,\cdot)$ is a functional in $\ell_2^m$, so there exists a representing vector $\boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}} \in \ell_2^m$ such that the $k$-th information value (as a random variable for fixed $\mathbf{y}_{[k-1]}$) is $L_{k,\mathbf{y}_{[k-1]}}(J\mathbf{X}) = \langle \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}}, \mathbf{X} \rangle$. Not waisting any information actually means that $L_{k,\mathbf{y}_{[k-1]}}(JP_{k-1,\mathbf{y}_{[k-2]}}\,\cdot)$ is not the zero functional, therefore we may assume $L_{k,\mathbf{y}_{[k-1]}}(JP_{k-1,\mathbf{y}_{[k-2]}}\,\cdot) = L_{k,\mathbf{y}_{[k-1]}}(J\,\cdot)$. This is equivalent to $P_{k-1,\mathbf{y}_{[k-2]}}\boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}} = \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}}$, and by induction it further implies the orthogonality $L_{k,\mathbf{y}_{[k-1]}}(m_{\mathbf{y}_{[k-1]}}) = \langle \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}}, \mathbf{n}_{\mathbf{y}_{[k-1]}} \rangle = 0$. In addition, we may assume that $\|\boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}}\|_2 = 1$ such that $\langle \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}}, \mathbf{X} \rangle$ is a standard Gaussian random variable. We now set

$$P_{k,\mathbf{y}_{[k-1]}}\mathbf{x} := P_{k-1,\mathbf{y}_{[k-2]}}\mathbf{x} - \langle \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}}, \mathbf{x} \rangle \, \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}} \,, \quad \text{and} \quad \mathbf{n}_{\mathbf{y}_{[k]}} := \mathbf{n}_{\mathbf{y}_{[k-1]}} + y_k \, \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}} \,,$$

thus defining $\tilde{\mu}_{\mathbf{y}_{[k]}}$. Note that by construction $P_{k,\mathbf{y}_{[k-1]}}\mathbf{n}_{\mathbf{y}_{[k]}} = \mathbf{0}$. Then for any

[6] $\mathbb{R}^0 = \{0\}$ is the zero vector space.

measurable set $E \subseteq \widetilde{F}$ we have

$$\begin{aligned}
\tilde{\mu}_{\mathbf{y}_{[k-1]}}(E) &= \mathbb{P}\{JP_{k-1,\mathbf{y}_{[k-2]}}\mathbf{X} + m_{\mathbf{y}_{[k-1]}} \in E\} \\
&= \mathbb{P}\{J(P_{k,\mathbf{y}_{[k-1]}}\mathbf{X} + \langle \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}}, \mathbf{X}\rangle \, \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}}) + J\mathbf{n}_{\mathbf{y}_{[k-1]}} \in E\} \\
&\overset{(*)}{=} \int_{\mathbb{R}} \mathbb{P}\{JP_{k,\mathbf{y}_{[k-1]}}\mathbf{X} + \underbrace{J(\mathbf{n}_{\mathbf{y}_{[k-1]}} + y_k \, \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}})}_{=J\mathbf{n}_{\mathbf{y}_{[k]}}=m_{\mathbf{y}_{[k]}}} \in E\} \, \gamma_1(\mathrm{d}y_k) \\
&\overset{\text{def.}}{=} \int_{\mathbb{R}} \tilde{\mu}_{\mathbf{y}_{[k]}}(E) \, \gamma_1(\mathrm{d}y_k) \, .
\end{aligned}$$

The step $(*)$ of splitting the integral into two integrations was possible because

- the Gaussian random vector $P_{k,\mathbf{y}_{[k-1]}}\mathbf{X}$ is stochastically independent from the Gaussian random variable $\langle \boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}}, \mathbf{X}\rangle$ due to orthogonality, and
- the span of $J\boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}}$ is not inside the image of $JP_{k,\mathbf{y}_{[k-1]}}$, which provides that for any $f \in E \cap \operatorname{supp}(\tilde{\mu}_{\mathbf{y}_{[k-1]}}) = E \cap (\operatorname{img}(JP_{k-1,\mathbf{y}_{[k-2]}}) + m_{\mathbf{y}_{[k-1]}})$ there is a unique representation $f = f_k + y_k J\boldsymbol{\xi}_{k,\mathbf{y}_{[k-1]}} + m_{\mathbf{y}_{[k]}}$ with $y_k \in \mathbb{R}$ and a vector $f_k \in \operatorname{img}(JP_{k,\mathbf{y}_{[k-1]}})$.

Now, by induction, we have

$$\tilde{\mu}(E) = \int_{\mathbb{R}^{k-1}} \tilde{\mu}_{\mathbf{y}_{[k-1]}}(E) \, \gamma_{k-1}(\mathrm{d}\mathbf{y}_{[k-1]}) \, ,$$

which by the above results and the product structure of the standard Gaussian measure may be continued as

$$\begin{aligned}
&= \int_{\mathbb{R}^{k-1}} \int_{\mathbb{R}} \tilde{\mu}_{\mathbf{y}_{[k]}}(E) \, \gamma_1(\mathrm{d}y_k) \, \gamma_{k-1}(\mathrm{d}\mathbf{y}_{[k-1]}) \\
&= \int_{\mathbb{R}^k} \tilde{\mu}_{\mathbf{y}_{[k]}}(E) \, \gamma_k(\mathrm{d}\mathbf{y}_{[k]}) \, .
\end{aligned}$$

The lemma is obtained for $k = n$ with $P_{\mathbf{y}} = P_{n,\mathbf{y}_{[n-1]}}$ and $m_{\mathbf{y}} = J\mathbf{n}_{\mathbf{y}_{[n]}}$. □

Having the representation of the conditional measure for the untruncated Gaussian measure, it is of interest to know the probability of obtaining information such that the mass of the conditional measure is concentrated inside the unit ball $F \subseteq \widetilde{F}$ and therefore truncation is not a great loss.

**Lemma 2.3.** *In the situation of Lemma 2.2, with* $\rho := \mathbb{E}\,\|J\mathbf{X}\|_F / \|J\|_{2\to F}$, *and the image measure* $\tilde{\nu} := \tilde{\mu} \circ N^{-1}$, *for* $\kappa > 1$ *we have the estimate*

$$\begin{aligned}
\tilde{\nu}\{\mathbf{y} \,:\, \|m_{\mathbf{y}}\|_F \leq 2\,\kappa \, \mathbb{E}\,\|J\mathbf{X}\|_F\} &\geq 1 - 2\,\exp\left(-\frac{(\kappa-1)^2\,\rho^2}{2}\right) \\
&\geq 1 - 2\,\exp\left(-\frac{(\kappa-1)^2}{\pi}\right) .
\end{aligned}$$

*Proof.* Writing $t := \kappa\, \mathbb{E}\, \|J\mathbf{X}\|_F$, basic estimates give

$$\begin{aligned}
\tilde{\nu}\{\mathbf{y} \,:\, \|m_{\mathbf{y}}\|_F \le 2\,t\} &\ge \tilde{\mu}\{f \,:\, \|f\|_F \le t\} \\
&\quad - \tilde{\mu}\{f \,:\, \|m_{\mathbf{y}}\|_F > 2\,t \text{ with } \mathbf{y} = N(f) \text{ and } \|f\|_F \le t\} \\
&\ge 1 - \tilde{\mu}\{f \,:\, \|f\|_F > t\} - \sup_{\mathbf{y}} \tilde{\mu}_{\mathbf{y}}\{f \,:\, \|f - m_{\mathbf{y}}\|_F > t\} \\
&\ge 1 - \mathbb{P}\{\|J\mathbf{X}\|_F > t\} - \sup_{\substack{P \text{ orth. proj.} \\ \operatorname{rk} P = m-n}} \mathbb{P}\{\|JP\mathbf{X}\|_F \ge t\}\,.
\end{aligned}$$

Applying Corollary A.6 with $t > \mathbb{E}\, \|J\mathbf{X}\|_F$ gives us

$$\ge 1 - \exp\left(-\frac{(t - \mathbb{E}\, \|J\mathbf{X}\|_F)^2}{2\, \|J\|_{2\to F}^2}\right) - \exp\left(-\frac{(t - \mathbb{E}\, \|J\mathbf{X}\|_F)^2}{2\, \|JP\|_{2\to F}^2}\right),$$

which by $\mathbb{E}\, \|JP\mathbf{X}\|_F \le \mathbb{E}\, \|J\mathbf{X}\|_F$, see Lemma A.3, and $\|JP\|_{2\to F} \le \|J\|_{2\to F}$, reduces to

$$\begin{aligned}
&\ge 1 - 2\, \exp\left(-\frac{(t - \mathbb{E}\, \|J\mathbf{X}\|_F)^2}{2\, \|J\|_{2\to F}^2}\right) \\
&= 1 - 2\, \exp\left(-\frac{(\kappa - 1)^2\, \rho^2}{2}\right).
\end{aligned}$$

The first lower bound is meaningful for $\kappa > 1 + \sqrt{2 \log 2}/\rho$, otherwise it is not positive and should be replaced by the trivial lower bound 0.

Lemma A.4 gives us the general estimate $\mathbb{E}\, \|J\mathbf{X}\|_F / \|J\|_{2\to F} \ge \sqrt{2/\pi}$, which leads to the second lower bound. This now is meaningful for $\kappa > 1 + \sqrt{\pi \log 2}$. □

**Example 2.4** (Why $\|m_{\mathbf{y}}\|_F \le r$ is a complicated constraint)**.** We consider $\widetilde{F} = \ell_\infty^m$ for $m \ge 2$ and $J = \mathrm{id} : \ell_2^m \to \ell_\infty^m$. The center $m_{\mathbf{y}} = \mathbf{n}_{\mathbf{y}}$ of the conditional distribution on the affine subspace $N^{-1}(\mathbf{y}) \subset \widetilde{F}$ of inputs $f = \mathbf{x}$ with the same information $\mathbf{y}$ is orthogonal to that subspace, i.e. $\mathbf{n}_{\mathbf{y}} \perp (\mathbf{x} - \mathbf{n}_{\mathbf{y}})$ for all $\mathbf{x} \in N^{-1}(\mathbf{y})$.

Consider the situation

$$m_{\mathbf{y}} = \mathbf{n}_{\mathbf{y}} = \left(1, \tfrac{1}{\sqrt{m+1}}, \ldots, \tfrac{1}{\sqrt{m+1}}\right) \in \mathbb{R}^m\,.$$

If, for example, $N(f) := \langle \mathbf{n}_{\mathbf{y}}, f \rangle \in \mathbb{R}^1$ then

$$f = \mathbf{x} = \left(\tfrac{2}{\sqrt{m+1}}, \ldots, \tfrac{2}{\sqrt{m+1}}\right) \in \mathbb{R}^m$$

lies within $N^{-1}(\mathbf{y})$ because

$$(\mathbf{x} - \mathbf{n}_{\mathbf{y}}) \perp \mathbf{n}_{\mathbf{y}}\,.$$

In this situation we have

$$\|m_{\mathbf{y}}\|_\infty = 1 \quad \text{and} \quad \|f\|_\infty = \tfrac{2}{\sqrt{m+1}} \xrightarrow[m\to\infty]{} 0\,.$$

### 2.2.2 Norm Expectation of Truncated Gaussian Measures

The following lemma is a simplification of a result in Heinrich [22, Lem 1]. This simplification is only feasible in the situation of Bernstein numbers. The more complicated version is given in Lemma 2.6.

**Lemma 2.5.** *For $J : \ell_2^m \to \widetilde{F}$ and $\mathbf{X}$ being the standard Gaussian vector in $\mathbb{R}^m = \ell_2^m$, set $\rho := \mathbb{E}\,\|J\mathbf{X}\|_F / \|J\|_{2\to F}$. Then for $\lambda > 1$ we have*

$$\begin{aligned}
&\mathbb{E}\Big[\|J\mathbf{X}\|_F \mathbb{1}_{\{\|J\mathbf{X}\|_F \le \lambda\, \mathbb{E}\,\|J\mathbf{X}\|_F\}}\Big] \\
&\qquad \ge \underbrace{\left[1 - \left(\lambda + \frac{1}{(\lambda-1)\,\rho^2}\right) \exp\left(-\frac{(\lambda-1)^2\,\rho^2}{2}\right)\right]_+}_{=:\beta(\lambda,\rho)} \mathbb{E}\,\|J\mathbf{X}\|_F \\
&\qquad\qquad \ge \underbrace{\left[1 - \left(\lambda + \frac{\pi}{2\,(\lambda-1)}\right) \exp\left(-\frac{(\lambda-1)^2}{\pi}\right)\right]_+}_{=:\beta(\lambda)} \mathbb{E}\,\|J\mathbf{X}\|_F\,.
\end{aligned}$$

(Note that $\beta(\lambda)$ vanishes for $\lambda \le 3.0513$, but it is positive for $\lambda \ge 3.0514$ and monotonically increasing with limit $\beta(\lambda) \xrightarrow[\lambda\to\infty]{} 1$.)

*Proof.* With $t := \lambda\, \mathbb{E}\,\|J\mathbf{X}\|_F$ we have

$$\mathbb{E}\Big[\|J\mathbf{X}\|_F \mathbb{1}_{\{\|J\mathbf{X}\|_F \le t\}}\Big] = \mathbb{E}\,\|J\mathbf{X}\|_F - t\,\mathbb{P}\{\|J\mathbf{X}\|_F \ge t\} - \int_t^\infty \mathbb{P}\{\|J\mathbf{X}\|_F \ge s\}\,\mathrm{d}s\,.$$

By the deviation result Corollary A.6, and substituting $s = \kappa\, \mathbb{E}\,\|J\mathbf{X}\|_F$, we can bound this by

$$\ge \left[1 - \lambda \exp\left(-\frac{(1-\lambda)^2\,\rho^2}{2}\right) - \int_\lambda^\infty \exp\left(-\frac{(1-\kappa)^2\,\rho^2}{2}\right) \mathrm{d}\kappa\right] \mathbb{E}\,\|J\mathbf{X}\|_F\,.$$

Using the estimate

$$\begin{aligned}
\int_\lambda^\infty \exp\left(-\frac{(1-\kappa)^2\,\rho^2}{2}\right) \mathrm{d}\kappa &\le \int_\lambda^\infty \frac{\kappa-1}{\lambda-1} \exp\left(-\frac{(\kappa-1)^2\,\rho^2}{2}\right) \mathrm{d}s \\
&= \frac{1}{(\lambda-1)\,\rho^2} \exp\left(-\frac{(\lambda-1)^2\,\rho^2}{2}\right),
\end{aligned}$$

we obtain the final lower bound.

The factor $\beta(\lambda,\rho)$ is monotonically increasing in $\rho$, so taking the general bound $\rho \ge \sqrt{2/\pi}$, see Lemma A.4, we obtain the second estimate.

Of course, the truncated expectation is non-negative, so we take the positive part $[\ldots]_+$ of the prefactor. □

For comparison, we cite the original lemma from Heinrich [22, Lem 1] concerning the truncated norm expectation when dealing with two different norms at once.

**Lemma 2.6.** *Consider a similar situation to Lemma 2.5 above with the ratios* $\rho := \mathbb{E}\,\|J\mathbf{X}\|_F / \|J\|_{2\to F}$ *and* $\sigma := \mathbb{E}\,\|SJ\mathbf{X}\|_G / \|SJ\|_{2\to G}$. *Then for* $\kappa, \lambda > 1$ *we have*

$$\begin{aligned}
&\mathbb{E}\Big[\|SJ\mathbf{X}\|_G \mathbb{1}_{\{\|J\mathbf{X}\|_F \le \kappa\, \mathbb{E}\,\|J\mathbf{X}\|_F\}}\Big] \\
&\qquad \ge \underbrace{\left[\beta(\lambda,\sigma) - \lambda \exp\left(-\frac{(\kappa-1)^2\,\rho^2}{2}\right)\right]_+}_{=:\tilde\beta(\kappa,\lambda,\rho,\sigma)} \mathbb{E}\,\|SJ\mathbf{X}\|_F \\
&\qquad\qquad \ge \underbrace{\left[\beta(\lambda) - \lambda \exp\left(-\frac{(\kappa-1)^2}{\pi}\right)\right]_+}_{=:\tilde\beta(\kappa,\lambda)} \mathbb{E}\,\|SJ\mathbf{X}\|_F\,.
\end{aligned}$$

*Idea of the proof.* The trick is that we replace the truncation with respect to the $F$-norm by a truncation with respect to the $G$-norm, for that purpose introducing an auxiliary parameter $\lambda$. We estimate the difference between both truncation variants,

$$\begin{aligned}
&\mathbb{E}\Big[\|SJ\mathbf{X}\|_G \mathbb{1}_{\{\|J\mathbf{X}\|_F \le \kappa\, \mathbb{E}\,\|J\mathbf{X}\|_F\}}\Big] \\
&\ge \mathbb{E}\Big[\|SJ\mathbf{X}\|_G \mathbb{1}_{\{\|SJ\mathbf{X}\|_G \le \lambda\, \mathbb{E}\,\|SJ\mathbf{X}\|_G\}}\Big] - (\lambda\, \mathbb{E}\,\|SJ\mathbf{X}\|_G)\, \mathbb{P}\{\|J\mathbf{X}\|_F > \kappa\, \mathbb{E}\,\|J\mathbf{X}\|_F\}\,.
\end{aligned}$$

The first term may be estimated by applying Lemma 2.5 to the operator $SJ$, for the second term we can directly use the deviation result Corollary A.6. □

**Remark 2.7** (Heinrich's original lower bound)**.** Heinrich's result [22, Prop 2] originally provides lower bounds for the Monte Carlo error via norm expectations of Gaussian measures. In detail, there exists a constant $c' > 0$ such that for $m > n$ and any injective linear operator $J : \ell_2^m \to \widetilde{F}$ we have

$$e^{\mathrm{ran,ada}}(n, S, F, \Lambda^{\mathrm{all}}) \ge c' \inf_{\substack{P \text{ orth. Proj.} \\ \mathrm{rk}\, P = m-n}} \frac{\mathbb{E}\,\|SJP\mathbf{X}\|_G}{\mathbb{E}\,\|J\mathbf{X}\|_F}\,, \tag{2.2.3}$$

where $\mathbf{X}$ is a standard Gaussian random vector in $\mathbb{R}^m = \ell_2^m$.

The proof works similarly to the proof of Theorem 2.1. In detail, for any $\alpha > 0$ one may rescale the operator $J$ such that $\|J\mathbf{X}\|_F \overset{!}{=} \alpha$, we truncate the rescaled measure. The constant then is determined as

$$c' = \nu(r,\alpha)\, \tilde\beta\left(\tfrac{r}{\alpha}, \lambda\right)\, \alpha\,,$$

now applying Lemma 2.6 instead of Lemma 2.5. With $r = 0.375$, $\alpha = 0.073$, and $\lambda = 6.15$, we obtain $c' = 0.06635...$ which is not much worse than the constant $c$ in Theorem 2.1.

How should $J$ be chosen? When applying Corollary 2.9 from the next section to find an optimal $J' := SJ$, that is, the image measure $\tilde\mu \circ S^{-1}$ shall be "well spread" within $G$, we may get rid of the infimum within (2.2.3) and write

$$e^{\mathrm{ran,ada}}(n, S, F, \Lambda^{\mathrm{all}}) \ge c'\, \frac{m-n}{m}\, \frac{\mathbb{E}\,\|SJ\mathbf{X}\|_G}{\mathbb{E}\,\|J\mathbf{X}\|_F}\,. \tag{2.2.4}$$

Especially for the identity mapping between sequence spaces $\ell_p^m \hookrightarrow \ell_q^m$, the optimal Gaussian measure will be the standard Gaussian measure, see Remark 2.10.

### 2.2.3 Optimal Gaussian Measures

#### Lewis' Theorem and Application to Gaussian Measures

We want to find optimal Gaussian measures with respect to the $F$-norm in $\mathbb{R}^m$. We therefore apply Lewis' Theorem, originally [42]. The proof given here is taken from Pisier [66, Thm 3.1]. It is included for completeness.

**Proposition 2.8** (Lewis' Theorem)**.** *Let $\alpha$ be an arbitrary norm on the space of automorphisms $\mathcal{L}(\mathbb{R}^m)$. Let $\widetilde{J} \in \mathcal{L}(\mathbb{R}^m)$ maximize the determinant $\det(J)$ subject to $\alpha(J) = 1$. Then for any operator $T \in \mathcal{L}(\mathbb{R}^m)$ we have*

$$\operatorname{tr}(\widetilde{J}^{-1}T) \le m\,\alpha(T)\,.$$

*Proof.* Since any invertible operator $J$ can be rescaled such that $\alpha(J) = 1$, there are admissible operators that fulfil $\det(J) > 0$. The constraint $\alpha(J) = 1$ defines a compact subset within the finite dimensional space $\mathcal{L}(\mathbb{R}^m)$, that is, the supremum

$$\sup_{\alpha(J)=1} \det(J) > 0$$

is attained. Let $\widetilde{J}$ be a maximizer. Then for any $T \in \mathcal{L}(\mathbb{R}^m)$ and $\varepsilon > 0$ we have

$$\det\left(\frac{\widetilde{J} + \varepsilon\, T}{\alpha(\widetilde{J} + \varepsilon\, T)}\right) \le \det(\widetilde{J})\,.$$

By homogeneity, and after dividing by $\det(\widetilde{J})$,

$$\det(1 + \varepsilon\, \widetilde{J}^{-1}T) \le \Big(\alpha(\widetilde{J} + \varepsilon\, T)\Big)^m \overset{\Delta\text{-ineq.}}{\le} (1 + \varepsilon\,\alpha(T))^m\,.$$

Finally,

$$\operatorname{tr}(\widetilde{J}^{-1}T) = \lim_{\varepsilon\to 0} \frac{\det(1 + \varepsilon\, \widetilde{J}^{-1}T) - 1}{\varepsilon} \le \lim_{\varepsilon\to 0} \frac{(1 + \varepsilon\,\alpha(T))^m - 1}{\varepsilon} = m\,\alpha(T)\,.$$

□

**Corollary 2.9** (Optimal Gaussian measures)**.** *For any norm $\|\cdot\|_F$ on $\mathbb{R}^m$ there is an operator $J \in \mathcal{L}(\mathbb{R}^m)$ with $\mathbb{E}\,\|J\mathbf{X}\|_F = 1$ and*

$$\mathbb{E}\,\|JP\mathbf{X}\|_F \ge \frac{\operatorname{rk} P}{m}$$

*for any projection $P \in \mathcal{L}(\mathbb{R}^m)$. Here, $\mathbf{X}$ is a standard Gaussian vector in $\mathbb{R}^m$.*

*Proof.* First note that $\alpha(J) := \mathbb{E}\,\|J\mathbf{X}\|_F$ defines a norm on the space $\mathcal{L}(\mathbb{R}^m)$ of linear operators $J : \mathbb{R}^m \to \mathbb{R}^m$. Indeed, because the expectation operator $\mathbb{E}$ is linear and $\|J\cdot\|_F$ is a semi-norm for any linear operator $J$, and $\alpha(J) > 0$ if $J \neq 0$.

We then may apply Proposition 2.8 with $T = JP$ and $\operatorname{tr} P = \operatorname{rk} P$ for projections $P$.

□

### Properties and Examples of Optimal Gaussian Measures

The remaining part of this subsection collects some results that expand our knowledge on optimal Gaussian measures but are not necessary for the basic version of the chapter's main result, Theorem 2.1.

**Remark 2.10** (Uniqueness of the Gaussian measure)**.** For any orthogonal matrix $A \in \mathcal{L}(\mathbb{R}^m)$, the distribution of $A\mathbf{X}$ is identical to that of $\mathbf{X}$, of course, $|\det A| = 1$. Consequently, $\widetilde{J}$ and $\widetilde{J}A$ define the same distribution and are equivalent maximizers of the absolute value of the determinant $|\det J|$, subject to $\mathbb{E}\,\|J\mathbf{X}\|_F \overset{!}{=} 1$.

Moreover, $\widetilde{J}$ is unique modulo orthogonal transformations, i.e. for all similarly optimal operators $\widetilde{J}_1$ there exists an orthogonal matrix $A$ with $\widetilde{J}_1 = \widetilde{J}A$, see Pisier [66, Prop 3.6] for a proof. In particular, all similarly optimal operators have the same operator norm $\|\widetilde{J}_1\|_{2\to F} = \|\widetilde{J}\|_{2\to F}$, and there is one unique optimal Gaussian measure $\widetilde{\mu}$ associated with $F$.

Let us now consider sequence spaces $\ell_p^m$ with $1 \le p \le \infty$, and let $J\mathbf{X} \in \ell_p^m$ be a random vector distributed according to the corresponding optimal Gaussian measure. Due to the symmetry of sequence spaces, for any operator $Q$ permuting the coordinates of a vector in $\mathbb{R}^m$ and possibly changing some of their signs, clearly, the distribution of $QJ\mathbf{X}$ will be optimal as well. By the uniqueness we conclude that the covariance matrix must be a multiple of the identity, so the optimal Gaussian measure for sequence spaces is a scaled standard Gaussian vector. In other words, the optimal operator $\widetilde{J}$ may be chosen as a multiple of the identity, $\widetilde{J} = c\,\mathrm{id}_{\mathbb{R}^m}$.

Compare Mathé [46, Lem 4] for similar symmetry arguments in a more general setting. This has been used by Heinrich [22] to prove properties for standard Gaussian measures on sequences spaces $\ell_p^m$ which we obtain by optimality according to Lewis' theorem.

Next, we find bounds for the operator norm of an optimal operator corresponding to the optimal Gaussian measure on $\widetilde{F} = \mathbb{R}^m$. It is not known to the author whether this particular upper bound has already been proven before, however, its implication together with the deviation result Corollary A.6 in high dimensions is not surprising, and similar results are known under names such as *concentration phenomenon* and *thin shell estimates*.

**Proposition 2.11.** *There exists a constant $c > 0$ such that for any norm $\|\cdot\|_F$ on $\mathbb{R}^m$, $m \ge 2$, the operator $\widetilde{J}$ defined as in Corollary 2.9 is bounded by*

$$\|\widetilde{J}\|_{2\to F} \le c\,(\log m)^{-1/2}\,.$$

*On the other hand, we have the lower bound*

$$\sqrt{\frac{\pi}{2}}\,m^{-1} \le \|\widetilde{J}\|_{2\to F}\,.$$

*Proof.* The lower bound is rather simple. Let $P_i$ denote the projections onto the $i$-th coordinate. Then, using $\|\widetilde{J}P_i\|_{2\to F} \le \|\widetilde{J}\|_{2\to F}$, we have

$$1 = \mathbb{E}\,\|\widetilde{J}\mathbf{X}\|_F \overset{\Delta\text{-ineq.}}{\le} \sum_{i=1}^{m} \mathbb{E}\,\|\widetilde{J}P_i\mathbf{X}\|_F \le m\sqrt{\frac{2}{\pi}}\,\|\widetilde{J}\|_{2\to F}\,.$$

Now for the upper bounds. There exists an orthogonal rank-1 projection $P_1$ on $\mathbb{R}^m = \ell_2^m$ such that

$$L := \|\widetilde{J}\|_{2\to F} = \|\widetilde{J}P_1\|_{2\to F}\,.$$

For the complementary rank-$(m-1)$ projection $P_2 := \mathrm{id} - P_1$ we have

$$\|\widetilde{J}P_2\|_{2\to F} \le \|\widetilde{J}\|_{2\to F} = L\,.$$

Due to orthogonality, we can split $\widetilde{J}\mathbf{X} = \widetilde{J}P_1\mathbf{X} + \widetilde{J}P_2\mathbf{X}$ into two independent zero-mean random vectors. Let $\beta := \mathbb{E}\,\|\widetilde{J}P_2\mathbf{X}\|_F$ denote the expectation of the norm of the second part. Clearly, by Corollary 2.9 we have

$$1 - \frac{1}{m} \le \beta \le 1\,. \tag{2.2.5}$$

Note that $\|\widetilde{J}P_1\mathbf{X}\|_F$ is a real random variable with probability density

$$p_1(t) := \sqrt{\frac{2}{\pi L^2}}\,\exp\left(-\frac{t^2}{2L^2}\right)\,\mathbb{1}[t \ge 0]\,.$$

For the cumulative distribution function of the real random variable $\|\widetilde{J}P_2\mathbf{X}\|_F$ we write $F_2(s) := \mathbb{P}\{\|\widetilde{J}P_2\mathbf{X}\|_F \le s\}$, and by $p_2(s) := \frac{\mathrm{d}}{\mathrm{d}s}F_2(s)$ we denote the corresponding density function for $s > 0$. Proposition A.5 directly implies that for $s > \beta$ we have

$$F_2(s) \ge 1 - \exp\left(-\frac{(s-\beta)^2}{2L^2}\right)\,. \tag{2.2.6}$$

Now, by symmetry and independence we have

$$\begin{aligned} 1 = \mathbb{E}\,\|\widetilde{J}\mathbf{X}\|_F &= \frac{1}{2}\,\mathbb{E}\left[\|\widetilde{J}P_1\mathbf{X} + \widetilde{J}P_2\mathbf{X}\|_F + \|\widetilde{J}P_1\mathbf{X} - \widetilde{J}P_2\mathbf{X}\|_F\right] \\ &\overset{\Delta\text{-ineq.}}{\ge} \mathbb{E}\max\left(\|\widetilde{J}P_1\mathbf{X}\|_F, \|\widetilde{J}P_2\mathbf{X}\|_F\right)\,. \end{aligned} \tag{2.2.7}$$

It follows

$$\begin{aligned} \frac{1}{m} \overset{(2.2.5)}{\ge} 1 - \beta &= \mathbb{E}\,\|\widetilde{J}\mathbf{X}\|_F - \mathbb{E}\,\|\widetilde{J}P_2\mathbf{X}\|_F \\ &\overset{(2.2.7)}{\ge} \mathbb{E}\left[\max\left(\|\widetilde{J}P_1\mathbf{X}\|_F, \|\widetilde{J}P_2\mathbf{X}\|_F\right) - \|\widetilde{J}P_2\mathbf{X}\|_F\right] \\ &= \mathbb{E}[\underbrace{(\|\widetilde{J}P_1\mathbf{X}\|_F - \|\widetilde{J}P_2\mathbf{X}\|_F)}_{=:r}\,\mathbb{1}_{\{\|\widetilde{J}P_1\mathbf{X}\|_F \ge \|\widetilde{J}P_2\mathbf{X}\|_F\}}] \\ &= \int_0^\infty r \int_0^\infty p_1(r+s)p_2(s)\,\mathrm{d}s\,\mathrm{d}r \\ &\overset{\text{part. int.}}{=} \int_0^\infty r\left[\underbrace{p_1(r+s)F_2(s)\Big|_0^\infty}_{=0} - \int_0^\infty \underbrace{p_1'(r+s)}_{<0}\,F_2(s)\,\mathrm{d}s\right]\mathrm{d}r \\ &\overset{(2.2.6)}{\ge} \int_0^\infty r\left[-\int_0^\infty p_1'(r+t+\beta)\left(1 - \exp\left(-\frac{t^2}{2L^2}\right)\right)\,\mathrm{d}t\right]\mathrm{d}r \end{aligned}$$

$$\overset{\text{part. int.}}{=} \int_0^\infty r \Bigg[ \underbrace{- p_1(r+t+\beta)\left(1-\exp\left(-\frac{t^2}{2L^2}\right)\right)\Big|_0^\infty}_{=0} + L^{-2}\int_0^\infty t\, p_1(r+t+\beta)\,\exp\left(-\frac{t^2}{2L^2}\right)\,\mathrm{d}t\Bigg]\,\mathrm{d}r$$
$$= \sqrt{\frac{2}{\pi}}\,L^{-3}\int_0^\infty r\,\exp\left(-\frac{(r+\beta)^2}{4L^2}\right)\int_0^\infty t\,\exp\left(-\frac{\left(t+\frac{r+\beta}{2}\right)^2}{L^2}\right)\,\mathrm{d}t\,\mathrm{d}r\,.$$

Note that $t \geq \frac{1}{2}\left(t+\frac{r+\beta}{2}\right)$ for $t \geq \frac{r+\beta}{2}$. After the substitution $\tau = t + \frac{r+\beta}{2}$, the inequality can be continued as

$$\geq \sqrt{\frac{2}{\pi}}\,\frac{1}{2}\,L^{-3}\int_0^\infty r\,\exp\left(-\frac{(r+\beta)^2}{4L^2}\right)\int_{r+\beta}^\infty \tau\,\exp\left(-\frac{\tau^2}{L^2}\right)\,\mathrm{d}\tau\,\mathrm{d}r$$
$$= \sqrt{\frac{2}{\pi}}\,\frac{1}{4}\,L^{-1}\int_0^\infty r\,\exp\left(-\frac{5(r+\beta)^2}{4L^2}\right)\,\mathrm{d}r\,.$$

Using $r \geq \frac{1}{2}(r+\beta)$ for $r \geq \beta$, and with the substitution $\rho = r+\beta$, we go on to

$$\geq \sqrt{\frac{2}{\pi}}\,\frac{1}{8}\,L^{-1}\int_{2\beta}^\infty \rho\,\exp\left(-\frac{5\rho^2}{4L^2}\right)\,\mathrm{d}\rho$$
$$\overset{\beta\leq 1}{\geq} \sqrt{\frac{2}{\pi}}\,\frac{1}{20}\,L\,\exp\left(-\frac{5}{L^2}\right)\,.$$

For $0 < L \leq \sqrt{2/\pi}$ we have

$$L \geq \sqrt{\frac{\pi}{2}}\,L^2 > \sqrt{\frac{\pi}{2}}\,20\,\frac{1}{1+\frac{20}{L^2}} > \sqrt{\frac{\pi}{2}}\,20\,\exp\left(-\frac{20}{L^2}\right)\,.$$

By this we finally obtain

$$\frac{1}{m} \geq \exp\left(-\frac{25}{L^2}\right)\,.$$

Inverting the inequality, we get

$$L \leq 5\,(\log m)^{-1/2}\,.$$

This is the proposition with $c = 5$. □

**Example 2.12.** As shown before in Remark 2.10, for sequence spaces $\ell_p^m$, due to symmetry, the operator $\widetilde{J}$ for the optimal Gaussian measure is a multiple of the identity.

Considering $\ell_1^m$, we observe

$$\mathbb{E}\,\|\mathbf{X}\|_1 = \mathbb{E}\sum_{i=1}^m |\mathbf{X}_i| = \sqrt{\frac{2}{\pi}}\,m\,.$$

Therefore, the choice $\widetilde{J} := \sqrt{\pi/2}\, m^{-1}\, \mathrm{id}_{\mathbb{R}^m}$ is optimal. The norm of $\widetilde{J}$ matches the lower bound in Proposition 2.11. Furthermore, for projections $P$ onto coordinates, that is, $P\mathbf{x} = \sum_{i\in I} x_i \mathbf{e}_i$ with an index set $I \subseteq \{1,\dots,m\}$, we have $\operatorname{rk} P = \#I$ and

$$\mathbb{E}\,\|\widetilde{J}P\mathbf{X}\|_1 = \frac{\operatorname{rk} P}{m}\,.$$

Therefore, the lower bound in Corollary 2.9 is sharp.

Now consider $\ell_\infty^m$. Clearly,

$$\|\,\mathrm{id}_{\mathbb{R}^m}\,\|_{2\to\infty} = 1\,.$$

Furthermore, we can show

$$\mathbb{E}\,\|\mathbf{X}\|_\infty \le C\sqrt{1+\log m}\,, \tag{2.2.8}$$

see Lemma A.9. We rescale with $\alpha = \alpha(m) := (\mathbb{E}\,\|\mathbf{X}\|_\infty)^{-1} \ge \frac{1}{C}(1+\log m)^{-1/2}$, i.e. $\widetilde{J} := \alpha\ \mathrm{id}_{\mathbb{R}^m}$ generates the optimal Gaussian measure, and the order of $\|\widetilde{J}\|_{2\to\infty} = \alpha(m)$ is determined precisely thanks to the upper bound from Proposition 2.11.

**Remark 2.13** (Improved constant for large $n$ and $m$)**.** Using Proposition 2.11, the constant $\frac{1}{15}$ in Theorem 2.1 can be improved for (extremely) large $n$ and $m$. In detail, we have constants $0 < c_m \xrightarrow[m\to\infty]{} \frac{1}{3}$ such that for $m \ge 2n$ we can state

$$e^{\mathrm{ran,ada}}(n, S, F, \Lambda^{\mathrm{all}}) \ge c_m\, \frac{m-n}{m}\, b_m(S)\,.$$

Following the proof of Theorem 2.1, thereby setting $r = \frac{1}{3}$ and $\alpha := \mathbb{E}\,\|J\mathbf{X}\|_F = \frac{r}{1+\delta}$ with $\delta = \delta(\rho)$, we attain an estimate with a factor

$$\tilde{\nu}\{\|m_{\mathbf{y}} \le \tfrac{2}{3}\}\ \underbrace{\inf_{\substack{P \text{ orth. Proj.}\\ \operatorname{rk} P = m-n}} \mathbb{E}[\|JP\mathbf{X}\|_F\, \mathbb{1}_{\{\|JP\mathbf{X}\|_F \le \frac{1}{3}\}}]}_{\ge\, \beta(\lambda,\rho/2) \text{ for } m\ge 2n}\ \alpha\,.$$

Using Lemma 2.3 and Lemma 2.5 with $\|JP\mathbf{X}\|_F / \|JP\|_{2\to F} \ge \rho/2$, we can choose a $\rho$-dependent constant

$$c = c(\rho) = \left[1 - 2\exp\left(-\frac{\delta^2\rho^2}{2}\right)\right]\left[1 - \left(1+\delta+\frac{4}{\delta\,\rho^2}\right)\exp\left(-\frac{\delta^2\rho^2}{8}\right)\right]\frac{1}{3\,(1+\delta)}\,.$$

This expression is monotonically growing in $\rho$, and converging to $\frac{1}{3}$ for $\rho\to\infty$, if we choose $\frac{1}{\rho} \prec \delta(\rho) \prec 1$. For example, with $\delta(\rho) = \delta_0/\sqrt{\rho}$ we have

$$c(\rho) = \left[1 - 2\exp\left(-\frac{\delta_0^2\,\rho}{2}\right)\right]\left[1 - \left(1+\frac{\delta_0}{\sqrt{\rho}}+\frac{4}{\delta_0\,\rho^{3/2}}\right)\exp\left(-\frac{\delta_0^2\,\rho}{8}\right)\right]\frac{1}{3\,(1+\frac{\delta_0}{\sqrt{\rho}})}\,.$$

Since by Proposition 2.11 we know $\rho = \rho_m \ge \sqrt{\log m}/5 \xrightarrow[m\to\infty]{} \infty$, this shows that we can choose $c_m := c(\rho_m) \xrightarrow[m\to\infty]{} \frac{1}{3}$. However, this convergence is extremely slow. If one is really interested in better constants, it is recommendable to include best knowledge about $\rho$ directly into the estimate.

## 2.3 Special Settings

We study two interesting modifications of the main result Theorem 2.1. They are non-essential for the applications in Section 2.4. First, in Section 2.3.1 we widen the class of admissible algorithms, now allowing varying cardinality $n(\omega, f)$. Still, we can show a similar inequality, however with worse constants. Second, in Section 2.3.2 we restrict to homogeneous algorithms and obtain an estimate with *sharp constants*, even for varying cardinality.

### 2.3.1 Varying Cardinality

Up to this point we ignored the additional feature of varying cardinality because it does not change the big point but gives us unpleasant constants that detract from the main relation. However, for lower bounds it is of interest to assume the most general shape for algorithms.

In Heinrich [22] results were actually given for algorithms with non-adaptively varying cardinality $n(\omega)$. In this setting, by Lemma 1.4 and Theorem 2.1, for $n \in \mathbb{N}$ we directly obtain an estimate like

$$\bar{e}^{\mathrm{ran}}(n, S) \geq \tfrac{1}{2}\, e^{\mathrm{ran}}(2n, S) > \tfrac{1}{60}\, b_{4n}(S)$$

We can even consider adaptively varying cardinality $n(\omega, f)$, and still get similar bounds.

**Theorem 2.14.** *Let $S : \widetilde{F} \to G$ be a compact linear operator between Banach spaces, and the input set $F$ be the unit ball in $\widetilde{F}$. Considering algorithms with adaptively varying cardinality $n(\omega, f)$, for $n \in \mathbb{N}$ we have*

$$\bar{e}^{\mathrm{ran,ada}}(n, S, F, \Lambda^{\mathrm{all}}) > \tfrac{1}{63}\, b_{4n}(S)\,.$$

*More generally, for any injective linear operator $J : \ell_2^{8n} \to \widetilde{F}$, we can estimate*

$$e^{\mathrm{ran,hom}}(n, S, F, \Lambda^{\mathrm{all}}) > \frac{1}{64} \inf_{\substack{P \text{ orth. Proj.} \\ \mathrm{rk}\, P = 4n}} \frac{\mathbb{E}\, \|SJP\mathbf{X}\|_G}{\mathbb{E}\, \|J\mathbf{X}\|_F}\,, \tag{2.3.1}$$

*where $\mathbf{X}$ is a standard Gaussian random vector in $\mathbb{R}^{8n} = \ell_2^{8n}$.*

*Proof.* The proof works similar to that of Theorem 2.1. Again, we assume $\widetilde{F} = \mathbb{R}^m$. As before, we define a measure $\tilde{\mu}$ as the distribution of $J\mathbf{X}$ with $\mathbf{X}$ being a standard Gaussian random vector in $\mathbb{R}^m = \ell_2^m$, and set $\mu$ to be the restriction of $\tilde{\mu}$ to the unit ball $F \subset \widetilde{F}$. We write $\alpha := \mathbb{E}\, \|J\mathbf{X}\|_F$.

In view of Lemma 1.5 (Bakhvalov's technique), we aim to bound the $\mu$-average error for a deterministic algorithm $\phi \circ N : \widetilde{F} \to G$ with varying cardinality $n(\mathbf{y}) = n(f)$ such that

$$\begin{aligned} \bar{n} &\geq \mathrm{card}(\phi \circ N, \mu) \\ &= \int n(f)\, \mu(\mathrm{d}f) \\ \text{[Markov's ineq.]} \quad &\geq 2\bar{n}\, \mu\{n(f) > 2\bar{n}\}\,. \end{aligned}$$

Thus, using the definition of the truncation, we can estimate

$$\begin{aligned}\tilde{\mu}\{n(f) \le 2\bar{n}\} &\ge \tilde{\mu}\{n(f) \le 2\bar{n} \text{ and } f \in F\} \\ &= \mu\{n(f) \le 2\bar{n}\} \\ &\ge \mu(F) - \frac{1}{2} \qquad (2.3.2) \\ \text{[Corollary A.6]} \quad &\ge \frac{1}{2} - \exp\left(-\left(\frac{1}{\alpha} - 1\right)^2 \Big/ \pi\right).\end{aligned}$$

The conditional measure $\tilde{\mu}_{\mathbf{y}}$ can be represented as the distribution of $JP_{\mathbf{y}}\mathbf{X} + m_{\mathbf{y}}$, with a suitable orthogonal projection $P_{\mathbf{y}}$ of rank $m - n(\mathbf{y})$, and a vector $m_{\mathbf{y}} \in F$, not very different from the case of fixed cardinality, compare Lemma 2.2. Again, we write $\tilde{\nu} := \tilde{\mu} \circ N^{-1}$ for the distribution of the information. Following the same arguments as in the proof of Theorem 2.1, and in addition restricting the integral to the case $n(\mathbf{y}) \le 2\bar{n}$, we obtain the estimate for $r \in (0,1)$

$$e(\phi \circ N, \mu) \ge \tilde{\nu}\{\mathbf{y} : n(\mathbf{y}) \le 2\bar{n} \text{ and } \|m_{\mathbf{y}}\|_F \le 1 - r\} \inf_{\substack{P \text{ orth. Proj.} \\ \operatorname{rk} P \ge m - 2\bar{n}}} \mathbb{E}\Big[\|JP\mathbf{X}\|_F \mathbb{1}_{\{\|JP\mathbf{X}\|_F \le r\}}\Big] b_m(S).$$

The first factor can be estimated using inequality (2.3.2) and a slight adjustment of Lemma 2.3,

$$\begin{aligned}&\tilde{\nu}\{\mathbf{y} : \|m_{\mathbf{y}}\|_F \le 1 - r\} \\ &\quad \ge \frac{1}{2} - \exp\left(-\left(\frac{1}{\alpha} - 1\right)^2 \Big/ \pi\right) - 2\exp\left(-\left(\frac{1-r}{2\,\alpha} - 1\right)^2 \Big/ \pi\right) \\ &\quad =: \bar{\nu}(r, \alpha).\end{aligned}$$

The second factor can be bounded from below by $\frac{1}{2}\,\beta(\frac{r}{\alpha})\,\alpha$, if the operator $J$ is chosen optimally. This estimate is exactly the same as in the proof of Theorem 2.1, here with $\frac{m-2\bar{n}}{m}$ instead of $\frac{m-n}{m}$, so

$$\bar{e}^{\text{avg,ada}}(\bar{n}, S, \mu) \ge \left(\frac{m - 2\bar{n}}{m}\right)_+ \bar{\nu}(r, \alpha)\,\beta\left(\frac{r}{\alpha}\right)\,\alpha\, b_m(S) =: \hat{\varepsilon}(\bar{n}).$$

This lower bound exhibits convexity, so by the subtle version of Lemma 1.5 (Bakhvalov's technique), taking $m = 4n$, we obtain

$$\bar{e}^{\text{ran,ada}}(n, S, F) \ge \underbrace{\tfrac{1}{2}\,\bar{\nu}(r, \alpha)\,\beta(\tfrac{r}{\alpha})\,\alpha}_{=:\bar{c}}\; b_{4n}(S).$$

With $r = 0.35$ and $\alpha = 0.07$, this gives us a constant $\bar{c} = 0.0159... > \frac{1}{63}$.

For the more general estimate with Gaussian measures, we take $m = 8n$ and the rough version of Lemma 1.5 (Bakhvalov's technique), inserting $\bar{n} = 2n$ in the adjusted version of the above estimates, and obtain a constant $\bar{c}' := \frac{1}{2}\,\bar{\nu}(r, \alpha)\,\tilde{\beta}(\frac{r}{\alpha}, \lambda)\,\alpha$. Choosing $r = 0.36$, $\alpha = 0.07$, and $\lambda = 6$, we get the numerical value $\bar{c}' = 0.0158... > \frac{1}{64}$. □

In regard of the estimate for fixed cardinality,

$$e^{\mathrm{ran}}(2n, S) > \tfrac{1}{30}\, b_{4n}(S)\,,$$

we see that taking twice as much information as in the varying cardinality setting gives us bigger lower bounds by roughly a factor two only.[7] However, several estimates involved in this proof seem to be far from optimal. For homogeneous algorithms, see Section 2.3.2, using Lemma 1.6 for the analysis of the average setting, we will obtain sharp estimates that equally hold for algorithms with fixed and varying cardinality. Here, we could not apply Lemma 1.6 because of the much more complicated situation arising from truncation. Anyway, even without the homogeneity assumption, we can state:

> If upper bounds achieved by Monte Carlo algorithms with fixed cardinality are close to the lower bounds obtained using Bernstein widths (or directly, Gaussian measures), *varying cardinality does not help a lot.*

### 2.3.2 Homogeneous Monte Carlo Methods

For linear problems (as considered within this chapter), common algorithms are homogeneous (and also non-adaptive).[8] There is a close and very basic connection to the normalized error.

**Lemma 2.15.** *Concerning the approximation of a compact linear operator $S : \widetilde{F} \to G$ between Banach spaces over the reals using homogeneous algorithms, the absolute error criterion with the* input set *$F \subset \widetilde{F}$ being the unit ball of $\widetilde{F}$ coincides with the normalized error criterion,*

$$e^{\star,\mathrm{hom}}(n, S, F, \Lambda) = e^{\star,\mathrm{hom}}_{\mathrm{normal}}(n, S, \widetilde{F} \setminus \{0\}, \Lambda)\,,$$

*where $\star \in \{\mathrm{det}, \mathrm{ran}\}$.*

*Proof.* If $A_n$ is a homogeneous algorithm that is defined for $f \in F$, it is naturally extended to $f \in \widetilde{F} \setminus \{0\}$ such that $A^\omega_n(f) = A^\omega_n(\frac{f}{\|f\|_F})$. Indeed, this is unproblematic since the information mapping as well is homogeneous. Then for $f \neq 0$ we have

$$\begin{aligned} e_{\mathrm{normal}}(A_n, f) &= \mathbb{E}\, \frac{\|S(f) - A^\omega_n(f)\|_G}{\|f\|_F} \\ &= \mathbb{E}\, \Big\| S(\tfrac{f}{\|f\|_F}) - A^\omega_n(\tfrac{f}{\|f\|_F}) \Big\|_G \\ &= e\left(A_n, \tfrac{f}{\|f\|_F}\right) \\ &\leq e(A_n, F)\,. \end{aligned}$$

[7] For the general estimate for Gaussian measures we lose roughly a factor 4, for both the error and the cardinality.

[8] For example for the identity on sequence spaces $\mathrm{id} : \ell^m_p \to \ell^m_q$, the basic structure of asymptotically best known algorithms is described in Section 3.2.2.

This proves "$\geq$".

Now, for any algorithm $A_n$, and any non-zero input $f \in F \setminus \{0\}$ from the unit ball, we have

$$\begin{aligned} e(A_n, f) &= \mathbb{E}\,\|S(f) - A_n^\omega(f)\|_G \\ &\leq \mathbb{E}\left[\frac{\|S(f) - A_n^\omega(f)\|_G}{\|f\|_F}\right] \\ &= e_{\mathrm{normal}}(A_n, f) \\ &\leq e_{\mathrm{normal}}(A_n, \widetilde{F} \setminus \{0\})\,. \end{aligned}$$

Trivially, $e(A_n, 0) = 0$, so this proves "$\leq$". □

**Theorem 2.16.** *For $S : \widetilde{F} \to G$ being a compact linear operator between Banach spaces, and the input set $F$ being the unit ball within $\widetilde{F}$, we have*

$$e^{\mathrm{ran,hom}}(n, S, F, \Lambda^{\mathrm{all}}) \geq \frac{m-n}{m}\, b_m(S) \quad \textit{for } m > n.$$

*In general, for any injective linear operator $J : \ell_2^m \to \widetilde{F}$ we have*

$$e^{\mathrm{ran,hom}}(n, S, F, \Lambda^{\mathrm{all}}) \geq \inf_{\substack{P \text{ orth. Proj.} \\ \mathrm{rk}\, P = m-n}} \frac{\mathbb{E}\,\|SJP\mathbf{X}\|_G}{\mathbb{E}\,\|J\mathbf{X}\|_F}\,, \tag{2.3.3}$$

*where $\mathbf{X}$ is a standard Gaussian random vector in $\mathbb{R}^m = \ell_2^m$.*

*Actually, the same Bernstein estimate holds for homogeneous algorithms with varying cardinality as well, for $\bar{n} \geq 0$ we can state*

$$\bar{e}^{\mathrm{ran,hom}}(\bar{n}, S, F, \Lambda^{\mathrm{all}}) \geq \left(\frac{m-\bar{n}}{m}\right)_+ b_m(S)\,.$$

*Proof.* Similarly to the proof of Theorem 2.1, we choose a Gaussian measure $\tilde{\mu}$ described as the distribution of $J\mathbf{X}$. Here, however, we take the scaling $\mathbb{E}\,\|J\mathbf{X}\|_F = 1$, so for measurable sets $E \subseteq \widetilde{F}$, by

$$\mu(E) := \int_E \|f\|_F\, \tilde{\mu}(\mathrm{d}f) = \mathbb{E}\,\|J\mathbf{X}\|_F\, \mathbb{1}_{\{J\mathbf{X} \in E\}}\,,$$

a probability measure on $\widetilde{F}$ is defined. By Lemma 2.15, and Bakhvalov's technique in the normalized error criterion setting (see Proposition 1.2 for a proof in the absolute error criterion setting), we have

$$e^{\mathrm{ran,hom}}(n, F) = e^{\mathrm{ran,hom}}_{\mathrm{normal}}(n, \widetilde{F} \setminus \{0\}) \geq e^{\mathrm{avg,hom}}_{\mathrm{normal}}(n, \mu)\,.$$

Now, for any homogeneous deterministic algorithm $A_n = \phi \circ N$, first with fixed cardinality, we have

$$\begin{aligned} e_{\mathrm{normal}}(A_n, \mu) &= \int \frac{\|S(f) - A_n(f)\|_G}{\|f\|_F}\, \mu(\mathrm{d}f) \\ &= \int \|S(f) - A_n(f)\|_G\, \tilde{\mu}(\mathrm{d}f) \\ &= \int_{\mathbb{R}^n} \left[\int \|S(f) - \phi(\mathbf{y})\|_G\, \tilde{\mu}_{\mathbf{y}}(\mathrm{d}f)\right] \tilde{\mu} \circ N^{-1}(\mathrm{d}\mathbf{y})\,. \end{aligned}$$

Using the representation of the conditional measure, see Lemma 2.2, and with the same symmetrization argument as in the proof of Theorem 2.1, we continue

$$\geq \inf_{\substack{P \text{ orth. Proj.} \\ \operatorname{rk} P = m-n}} \mathbb{E} \, \|SJP\mathbf{X}\|_G \, .$$

By the definition of Bernstein numbers for $\widetilde{F} = \mathbb{R}^m$, choosing an optimal $J$ as it is found in Corollary 2.9, we end up with

$$e_{\text{normal}}(A_n, \mu) \geq \frac{m-n}{m} \, b_m(S) \, .$$

Observe that the lower bound for the conditional error, given $\mathbf{y}$, exhibits the special structure of Lemma 1.6 with a convex function $\hat{\varepsilon}(\bar{n}) := \left(\frac{m-\bar{n}}{m}\right)_+ b_m(S)$. However, switching airily between different error criterions and measures, it is not immediate that this already implies lower bounds for homogeneous algorithms with varying cardinality, so we need to think about a modification of the proof of Lemma 1.5 that fits to the present situation. Let $A = (A^\omega)_{\omega \in \Omega}$ be a homogeneous Monte Carlo method with varying cardinality. Due to homogeneity, $n(\omega, \lambda f) = n(\omega, f)$ for $\lambda \in \mathbb{R} \setminus \{0\}$ and $f \in \widetilde{F}$, hence

$$\begin{aligned} \bar{n} := \sup_{f \in F} \mathbb{E} \, n(\omega, f) &= \sup_{f \in \widetilde{F}} \mathbb{E} \, n(\omega, f) \\ \text{[Fubini]} \quad &\geq \mathbb{E} \int n(\omega, f) \, \tilde{\mu}(\mathrm{d}f) \, . \end{aligned}$$

The key insight is that the error $e(A, F)$ can be related to the $\tilde{\mu}$-average setting, here we use that Lemma 2.15 holds for algorithms with varying cardinality as well,

$$\begin{aligned} e(A, F) &= e_{\text{normal}}(A, \widetilde{F} \setminus \{0\}) \\ \text{[Fubini]} \quad &\geq \mathbb{E} \int e_{\text{normal}}(A^\omega, f) \, \mu(\mathrm{d}f) \\ &= \mathbb{E} \int e(A^\omega, f) \, \tilde{\mu}(\mathrm{d}f) \, . \end{aligned}$$

For the inner integral we apply Lemma 1.6 to the $\tilde{\mu}$-average setting, for which we have the lower bound $\hat{\varepsilon}(n(\omega, \mathbf{y}))$ for the conditional average error, that is, with respect to $\tilde{\mu}_{\mathbf{y}}$. We obtain

$$\geq \mathbb{E} \, \hat{\varepsilon}(\operatorname{card}(A^\omega, \tilde{\mu})) \, .$$

From here we can proceed as in the last inequality chain within the proof of Lemma 1.5,

$$e(A, F) \geq \hat{\varepsilon}(\bar{n}) \, .$$

Hence the lower bound based on Bernstein numbers does even hold for algorithms with varying cardinality. For the direct estimate via general Gaussian measures, it depends on the particular situation how we can generalize the lower bound. □

**Remark 2.17.** The above theorem is optimal. Indeed, consider for example the identity $\mathrm{id}_{\ell_1^m} : \ell_1^m \hookrightarrow \ell_1^m$ with Bernstein number $b_m(\mathrm{id}_{\ell_1^m}) = 1$. Let $I = I(\omega) \subseteq \{1, \dots, m\}$ be a randomly chosen index set such that $\mathbb{P}\{i \in I\} = \frac{\bar{n}}{m}$, where $0 \le \bar{n} \le m$, and define the linear (homogeneous) Monte Carlo method

$$A_{\bar{n}}^{\omega}(\mathbf{x}) := \sum_{i \in I(\omega)} x_i \, \mathbf{e}_i \,, \quad \mathbf{x} \in \ell_1^m \,,$$

where $\mathbf{e}_i$ are the vectors of the standard basis. The cardinality is

$$\mathrm{card}(A) \;=\; \mathbb{E}\,\#I(\omega) \;=\; \sum_{i=1}^{m} \mathbb{P}\{i \in I\} \;=\; \bar{n} \,,$$

the error

$$e(A_{\bar{n}}, \mathrm{id}_{\ell_1^m}, \mathbf{x}) \;=\; \mathbb{E}\|\mathbf{x} - A_{\bar{n}}^{\omega}(\mathbf{x})\|_1 \;=\; \sum_{i=1}^{m} \mathbb{P}\{i \notin I(\omega)\}\, |x_i| \;=\; \tfrac{m-\bar{n}}{m}\, \|\mathbf{x}\|_1 \,,$$

so $e^{\mathrm{ran,hom}}(A_{\bar{n}}, \mathrm{id}_{\ell_1^m}) = \frac{m-\bar{n}}{m}$. On the other hand, by Theorem 2.16 we have the lower bound $\bar{e}^{\mathrm{ran,hom}}(\bar{n}, \mathrm{id}_{\ell_1^m}) \ge \frac{m-\bar{n}}{m}$. This shows that $A_{\bar{n}}$ is optimal. If $\bar{n} \in \mathbb{N}_0$, we can find a fixed-cardinality version for $A_{\bar{n}}$.

## 2.4 Applications

The first application on the recovery of sequences, Section 2.4.1, is meant to give a feeling for the potentials and limitations of Bernstein numbers when it comes to lower bounds for quantities from IBC. We also discuss other techniques for lower bounds, as well as general problems arising. The main application is the $L_\infty$-approximation of $C^\infty$-functions, see Section 2.4.2. There we show that in the particular situation randomization does not help in terms of tractability classifications.

### 2.4.1 Recovery of Sequences

We consider the approximation of the identity between finite dimensional sequence spaces,

$$\mathrm{APP} : \ell_p^M \hookrightarrow \ell_q^M,$$

with $1 \le p, q \le \infty$ and $M \in \mathbb{N}$. This example problem is also of interest when dealing with embeddings between function spaces, compare Section 3.2.3.

**The Case $M = 4n$**

The following result on Bernstein numbers is well known and has been used e.g. in [38, 49, 50] in order to determine the order of decay of Bernstein numbers in different function space settings,

$$b_m(\ell_p^{2m} \hookrightarrow \ell_q^{2m}) \asymp \begin{cases} m^{1/q-1/p} & \text{if } 1 \le p \le q \le \infty \text{ or } 1 \le q \le p \le 2 \,, \\ m^{1/q-1/2} & \text{if } 1 \le q \le 2 \le p \le \infty \,, \\ 1 & \text{if } 2 \le q \le p \le \infty \,. \end{cases}$$

Here, the hidden constants may depend on $p$ and $q$. Applying Theorem 2.1 with $m = 2n$, this implies the estimate

$$e^{\mathrm{ran}}(n, \ell_p^{4n} \hookrightarrow \ell_q^{4n}) \succeq \begin{cases} n^{1/q-1/p} & \text{if } 1 \le p \le q \le \infty \text{ or } 1 \le q \le p \le 2\,, \\ n^{1/q-1/2} & \text{if } 1 \le q \le 2 \le p \le \infty\,, \\ 1 & \text{if } 2 \le q \le p \le \infty\,. \end{cases}$$

Since the lower bounds with Bernstein numbers were obtained using Gaussian measures, it is not surprising that in some parameter settings we will get significantly better estimates when directly working with Gaussian measures, as it has been done in Heinrich [22], see Remark 2.7. In detail, with $\mathbf{X}$ being a standard Gaussian vector on $\mathbb{R}^{4n}$, by (2.2.4) we have[9]

$$e^{\mathrm{ran}}(n, \ell_p^{4n} \hookrightarrow \ell_q^{4n}) \ge c\, \frac{\mathbb{E}\,\|\mathbf{X}\|_q}{\mathbb{E}\,\|\mathbf{X}\|_p} \succeq \begin{cases} n^{1/q-1/p} & \text{if } 1 \le p, q < \infty\,, \\ n^{-1/p}(\log n)^{1/2} & \text{if } 1 \le p < q = \infty\,, \\ n^{1/q}(\log n)^{-1/2} & \text{if } 1 \le q < p = \infty\,, \\ 1 & \text{if } p = q = \infty\,, \end{cases} \tag{2.4.1}$$

where $c > 0$ is a universal constant, and the hidden constant for the second relation may depend on $p$ and $q$. For this result we only need to know $\mathbb{E}\,\|\mathbf{X}\|_p \asymp m^{1/p}$ for $1 \le p < \infty$, and $\mathbb{E}\,\|\mathbf{X}\|_\infty \asymp \sqrt{1+\log(m)}$, for a standard Gaussian vector $\mathbf{X}$ in $\mathbb{R}^m = \mathbb{R}^{4n}$, see Lemma A.9. In Heinrich [22] we also find upper bounds, which are achieved by non-adaptive and homogeneous methods, see also Section 3.2.2,

$$e^{\mathrm{ran}}(n, \ell_p^{4n} \hookrightarrow \ell_q^{4n}) \preceq \begin{cases} n^{1/q-1/p} & \text{if } 1 \le p, q < \infty\,, \\ n^{-1/p}(\log n)^{1/2} & \text{if } 1 \le p < q = \infty\,, \\ n^{1/q} & \text{if } 1 \le q < p = \infty\,, \\ 1 & \text{if } p = q = \infty\,. \end{cases} \tag{2.4.2}$$

We see that in most cases the lower bounds (2.4.1) obtained by Gaussian measures match the upper bounds (2.4.2), but for $q < p = \infty$ there is a logarithmic gap.

For non-adaptive algorithms this gap can be closed. Within the authors master's thesis, see also [39], Bernstein numbers have been related to the error of non-adaptive Monte Carlo methods by means of volume ratios.[10] In the general linear setting of this chapter, for $n < m$ we have

$$e^{\mathrm{ran,nonada}}(n, S) \ge \frac{m-n}{m+1} \sup_{X_m} \inf_{Y_{m-n}} \left( \frac{\mathrm{Vol}_{m-n}(S(F) \cap Y_{m-n})}{\mathrm{Vol}_{m-n}(B_G \cap Y_{m-n})} \right)^{1/(m-n)}, \tag{2.4.3}$$

where $X_m \subseteq \widetilde{F}$ and $Y_{m-n} \subseteq S(X_m)$ are subspaces with dimension $\dim(X_m) = m$ and $\dim(Y_{m-n}) = m - n$, furthermore, $B_G$ denotes the unit ball in $G$, and for each choice

[9] We chose $m = 4n$ for better comparison with the results that were obtained via Bernstein numbers. In this case however, $m = 2n$ would give the same asymptotics.

[10] Instead of truncated Gaussian measures, in [39] the average case for the uniform distribution on finite-dimensional sub-balls of the input set $F$ was considered.

of $Y_{m-n}$ the volume measure $\mathrm{Vol}_{m-n}$ may be any $(m-n)$-dimensional Lebesgue-like measure since we are only interested in the ratio of volumes. In the case of sequence spaces $\ell_p^m \hookrightarrow \ell_q^m$, the volume ratios could actually be computed in [39] (based on results from Meyer and Pajor [48]), and by that we obtained

$$e^{\mathrm{ran,nonada}}(n, S) \succeq n^{1/q-1/p} \tag{2.4.4}$$

for the whole parameter range $1 \le p, q \le \infty$, thus closing the gap for $q < p = \infty$. For $p < q = \infty$, in turn, Gaussian measures do a better job.

### The Case of Small Cardinality $n \ll M$

Up to this point the dimension of the vector spaces in consideration was a constant times the number of information. That setting is good enough when aiming for the order of convergence for function space embeddings. But what if the cardinality of the information is much smaller than the size of the sequence spaces?

For an example of rather disappointing lower bounds we restrict to the case

$$\ell_1^M \hookrightarrow \ell_2^M .$$

There is a well known result on deterministic algorithms,

$$e^{\mathrm{det}}(n, \ell_1^M \hookrightarrow \ell_2^M) \asymp \min\left\{1, \sqrt{\frac{1+\log\frac{M}{n}}{n}}\right\} \quad \text{for } n < M.$$

A proof based on compressive sensing can be found e.g. in Foucart and Rauhut [17, Chap 10], the idea goes back to Kashin 1977 [30] and Garnaev and Gluskin 1984 [20]. These errors are obtained with homogeneous and non-adaptive algorithms, see Section 3.2.2.

What do we know about lower bounds for the Monte Carlo error? We could use the Bernstein numbers for $n < m \le M$,

$$b_m(\ell_1^M \hookrightarrow \ell_2^M) = \frac{1}{\sqrt{m}},$$

see Pinkus [64, pp. 202–205]. Then by Theorem 2.1, for $M \ge 3n$ we obtain

$$e^{\mathrm{ran}}(n, \ell_1^M \hookrightarrow \ell_2^M) \ge \frac{1}{15} \max_{n<m\le M} \frac{m-n}{m} \frac{1}{\sqrt{m}} \overset{m=3n}{=} \frac{1}{15} \frac{2}{3\sqrt{3}} \frac{1}{\sqrt{n}}.$$

It is not reasonable that the size $M$ of the problem did not contribute at all to the error quantity, so this lower bound for the Monte Carlo error is seemingly not quite optimal. Directly considering Gaussian measures does not change the big point, as the next lemma shows.

**Lemma 2.18.** *Let $J : \ell_2^m \to \mathbb{R}^M$ be an injective linear operator and $\mathbf{X}$ be a standard Gaussian random vector in $\ell_2^m$. Then for $n < m$ we have*

$$\inf_{\substack{P \text{ orth. Proj.} \\ \mathrm{rk}\, P = m-n}} \frac{\mathbb{E}\,\|JP\mathbf{X}\|_2}{\mathbb{E}\,\|J\mathbf{X}\|_1} \le \sqrt{\frac{\pi}{2}} \frac{\sqrt{m-n}}{m} \le \sqrt{\frac{\pi}{2}} \frac{1}{2} \frac{1}{\sqrt{n}}.$$

*Proof.* Let $\mathbf{r}_1, \dots, \mathbf{r}_M \in \ell_2^m$ denote the rows, and $\mathbf{c}_1, \dots, \mathbf{c}_m \in \mathbb{R}^M$ the columns of $J$. Without loss of generality, the columns of $J$ are orthogonal, compare Remark 2.10.

Obviously, the term

$$\inf_{\substack{P \text{ orth. Proj.} \\ \operatorname{rk} P = m-n}} \mathbb{E}\,\|JP\mathbf{X}\|_2$$

does not change when transforming $J$ into $\widetilde{J} := Q_M J Q_m$, with $Q_M$ being an orthogonal matrix operating on $\ell_2^M$, and $Q_m$ respectively on $\ell_2^m$. For the denominator we have

$$\mathbb{E}\,\|J\mathbf{X}\|_1 = \sqrt{\frac{2}{\pi}} \sum_{i=1}^{M} \|\mathbf{r}_i\|_2. \tag{2.4.5}$$

What happens to this, when we rotate rows (thus performing a transformation that contributes to $Q_M$)? Consider the rotation of the $i_1$-th and the $i_2$-th rows, with $\|r_{i_1}\|_2 > \|r_{i_2}\|_2$ and $\langle \mathbf{r}_{i_1}, \mathbf{r}_{i_2} \rangle \neq 0$, defined by

$$\begin{aligned} \mathbf{r}_{i_1} &\mapsto \mathbf{r}'_{i_1} := \sqrt{\xi}\,\mathbf{r}_{i_1} + \sqrt{1-\xi}\,\mathbf{r}_{i_2}\,, \\ \mathbf{r}_{i_2} &\mapsto \mathbf{r}'_{i_2} := -\sqrt{1-\xi}\,\mathbf{r}_{i_1} + \sqrt{\xi}\,\mathbf{r}_{i_2} \end{aligned}$$

with $\xi \in [0,1]$. By construction,

$$\|\mathbf{r}_{i_1}\|_2^2 + \|\mathbf{r}_{i_2}\|_2^2 = \|\mathbf{r}'_{i_1}\|_2^2 + \|\mathbf{r}'_{i_2}\|_2^2\,. \tag{2.4.6}$$

Now, for the special choice

$$\xi := \frac{1}{2}\left(1 + \frac{\operatorname{sgn}\langle \mathbf{r}_{i_1}, \mathbf{r}_{i_2} \rangle}{\sqrt{1 + \frac{4\langle \mathbf{r}_{i_1}, \mathbf{r}_{i_2} \rangle^2}{\|\mathbf{r}_{i_1}\|_2^2 - \|\mathbf{r}_{i_2}\|_2^2}}}\right),$$

one can check that

$$\langle \mathbf{r}_{i_1}, \mathbf{r}_{i_2} \rangle = 0\,, \quad \text{and} \quad \|\mathbf{r}'_{i_1}\|_2 > \|\mathbf{r}_{i_1}\|_2 \geq \|\mathbf{r}_{i_2}\|_2 > \|\mathbf{r}'_{i_2}\|_2\,.$$

Together with (2.4.6), one can easily prove that

$$\|\mathbf{r}'_{i_1}\|_2 + \|\mathbf{r}'_{i_2}\|_2 < \|\mathbf{r}_{i_1}\|_2 + \|\mathbf{r}_{i_2}\|_2\,.$$

This means that by such transformations performed on $J$, the expression (2.4.5) will be reduced. Now, repeatedly performing such transformations, and permuting rows of $J$, one can find a matrix $J' = Q_M J$ with orthogonal rows $\mathbf{r}'_1, \dots, \mathbf{r}'_M \in \ell_2^m$ of descending norm, in particular $\mathbf{r}'_{m+1} = \dots = \mathbf{r}'_M = \mathbf{0}$, moreover, $\mathbb{E}\,\|J'\mathbf{X}\|_1 \leq \mathbb{E}\,\|J\mathbf{X}\|_1$. Since the columns of $J$ are orthogonal, so are the columns of $J'$. Hence we can find an orthogonal matrix $Q_m$ such that the only non-zero entries of the matrix $\widetilde{J} = J'Q_m$ lie on the diagonal, $j_{kk} = \lambda_k$, and $\lambda_1 \geq \dots \geq \lambda_m > 0$. Without loss of generality,

$$\sum_{k=1}^{m} \lambda_k = 1\,, \tag{2.4.7}$$

so

$$\mathbb{E}\,\|J'\mathbf{X}\|_1 = \mathbb{E}\,\|\widetilde{J}\mathbf{X}\|_1 = \sqrt{\frac{2}{\pi}}\,.$$

Now, consider the projection $P_n$ onto the last $m-n$ coordinates, i.e. the mapping $P_n\mathbf{x} := \sum_{k=n+1}^{m} x_k\,\mathbf{e}_k$ with $\mathbf{e}_k$ being the standard basis in $\ell_2^m$. Then

$$\mathbb{E}\,\|\widetilde{J}P_n\mathbf{X}\|_2 \le \sqrt{\mathbb{E}\,\|\widetilde{J}P_n\mathbf{X}\|_2^2} = \sqrt{\sum_{k=n+1}^{m}\lambda_k^2} \le \frac{1}{\sqrt{m-n}}\sum_{k=n+1}^{m}\lambda_k \le \frac{\sqrt{m-n}}{m}\,.$$

Here, we used the general inequality $\|\mathbf{x}\|_1 \le \sqrt{d}\,\|\mathbf{x}\|_2$ for $\mathbf{x}\in\mathbb{R}^d$ with $d=m-n$, and (2.4.7) together with the monotonicity of the coefficients $\lambda_k$.

Altogether we have

$$\inf_{\substack{P \text{ orth. Proj.}\\ \operatorname{rk} P = m-n}} \frac{\mathbb{E}\,\|JP\mathbf{X}\|_2}{\mathbb{E}\,\|J\mathbf{X}\|_1} \le \frac{\mathbb{E}\,\|\widetilde{J}P_n\mathbf{X}\|_2}{\mathbb{E}\,\|\widetilde{J}\mathbf{X}\|_1} \le \sqrt{\frac{\pi}{2}}\,\frac{\sqrt{m-n}}{m}\,.$$

This is maximized for $m=2n$. □

The situation for volume ratios (2.4.3) is the same, see [39, Sec 3.1] for further hints. This disappointing phenomenon is not new, Vybíral [76] showed that Gaussian measures, as well as uniform distributions, are not suitable in this and many other cases to obtain lower bounds that – up to a constant – match the upper bounds. In his paper on *best m-term approximation*[11] he basically shows that in those particular average case settings the initial error is already too small. He proposes other average case settings which work perfectly for best $m$-term approximation but are hard to use in the information based complexity framework.

There are also situations where the lower bounds perfectly reflect the situation for $n \ll M$, consider

$$\ell_p^M \hookrightarrow \ell_q^M$$

for the parameter range $1 \le q \le p \le \infty$. The initial error is

$$e(0, \ell_p^M \hookrightarrow \ell_q^M) = M^{1/q-1/p} \ge 1\,,$$

whereas the lower bounds by Gaussian measures (2.4.1) give us

$$e^{\text{ran,ada}}(n, \ell_p^M \hookrightarrow \ell_q^M) \succeq M^{1/q-1/p} \quad \text{for } n \le M/4$$

in the parameter range $1 \le q \le p < \infty$ or $q=p=\infty$,[12] where the hidden constant may depend on $p$ and $q$. In the case $1 \le q < p = \infty$, the lower bounds by Gaussian measures are worse by a logarithmic factor $(\log M)^{-1/2}$, but in the non-adaptive setting (2.4.4) we get rid of that term. What does this mean for our strategies to approximate this embedding for $q \le p$? Basically, we have the two alternatives of either

[11] The concept of best-$m$-term approximation is not covered by our framework of information-based complexity, but results on that topic often have implications on the performance of some types of algorithms. For instance, in the case $\ell_1^M \hookrightarrow \ell_2^M$, typical algorithms usually return vectors with only $n$ non-zero entries, compare Section 3.2.2.

[12] Bernstein numbers only give comparably satisfying lower bounds for $1 \le q \le p \le 2$ or $q=p$.

- taking no information $n = 0$ and accepting the initial error, or
- taking full information $n = M$, thus having no error at all.

This is a reasonable approach to the problem, because any choice $n \leq M/4$ will be insufficient if we want to reduce the initial error by a significant factor, and taking at most four times as much information than really necessary is no big deal.

### 2.4.2 $L_\infty$-Approximation for $C^\infty$-Functions

We consider the $L_\infty$-approximation for subclasses of $C^\infty$-functions defined on the $d$-dimensional unit cube $[0,1]^d$,

$$\mathrm{APP} : C^\infty([0,1]^d) \hookrightarrow L_\infty[0,1]^d .$$

The space $C^\infty([0,1]^d)$ has no natural norm and it will be crucial for tractability what input sets we choose. Novak and Woźniakowski [56] considered the input set

$$F^d := \{f \in C^\infty([0,1]^d) \mid \|D^{\boldsymbol{\alpha}} f\|_\infty \leq 1 \text{ for } \boldsymbol{\alpha} \in \mathbb{N}_0^d\} .$$

Here, $D^{\boldsymbol{\alpha}} f = \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d} f$ denotes the partial derivative of $f$ belonging to a multi-index $\boldsymbol{\alpha} \in \mathbb{N}_0^d$. In their study, Novak and Woźniakowski showed that with this input set the problem suffers from the curse of dimensionality for deterministic algorithms. Since the proof was based on the Bernstein numbers, thanks to Theorem 2.1, the curse of dimensionality extends to randomized algorithms.

We will cover this case within a slightly more general setting, considering the input sets

$$F_p^d := \{f \in C^\infty([0,1]^d) \mid \|\nabla_{\mathbf{v}_k} \cdots \nabla_{\mathbf{v}_1} f\|_\infty \leq |\mathbf{v}_1|_p \cdots |\mathbf{v}_k|_p \text{ for all } k \in \mathbb{N}_0,\, \mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^d\} ,$$

where $\nabla_{\mathbf{v}} f$ denotes the directional derivative along a vector $\mathbf{v} \in \mathbb{R}^d$, and we write $|\mathbf{v}|_p$ for the $p$-norm of $\mathbf{v} \in \ell_p^d$, $1 \leq p \leq \infty$. Note that, indeed, this is a generalization of the original problem since $F^d = F_1^d$. The set $F_p^d$ can be seen as the unit ball of the space

$$\widetilde{F}_p^d := \{f \in C^\infty([0,1]^d) \mid \|f\|_{F_p} < \infty\} , \tag{2.4.8}$$

equipped with the norm

$$\|f\|_{F_p} := \sup_{\substack{k \in \mathbb{N}_0 \\ \mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^d}} |\mathbf{v}_1|_p^{-1} \cdots |\mathbf{v}_k|_p^{-1} \, \|\nabla_{\mathbf{v}_k} \cdots \nabla_{\mathbf{v}_1} f\|_\infty . \tag{2.4.9}$$

First, we aim for lower bounds, to this end starting with the Bernstein numbers of the restricted operator $\mathrm{APP} : \widetilde{F}_p^d \hookrightarrow L_\infty[0,1]^d$. The proof follows the lines of Novak and Woźniakowski [56].

**Proposition 2.19.** *For $1 \leq p < \infty$ we have*

$$b_m(\widetilde{F}_p^d \hookrightarrow L_\infty) = 1 \quad \text{for } n \leq 2^{\lfloor \frac{d^{1/p}}{3} \rfloor}.$$

*In the case $p = 1$, this even holds for $n \leq 2^{\lfloor d/2 \rfloor}$.*

*Proof.* Note that $\|\cdot\|_{F_p} \geq \|\cdot\|_\infty$, and therefore $b_m(\widetilde{F}_p^d \hookrightarrow L_\infty) \leq 1$ for all $m \in \mathbb{N}$.

We set $r := \lceil 2\, d^{1-1/p} \rceil$ and $s := \lfloor d/r \rfloor \geq \lfloor \frac{d^{1/p}}{2+d^{-1}} \rfloor \geq \lfloor \frac{d^{1/p}}{3} \rfloor$. Consider the following linear subspace of $\widetilde{F}_p^d$,

$$V_p^d := \big\{ f \mid f(\mathbf{x}) = \sum_{\mathbf{i}\in\{0,1\}^s} a_{\mathbf{i}}\, (x_1 + \ldots + x_r)^{i_1} \cdots (x_{r(s-1)+1} + \ldots + x_{rs})^{i_s}, \quad a_{\mathbf{i}} \in \mathbb{R} \big\} \tag{2.4.10}$$

with $\dim V_p^d = 2^s$. For $f \in V_p^d$ and $\mathbf{v} \in \mathbb{R}^d$, we will show $\|\nabla_{\mathbf{v}} f\|_\infty \leq |\mathbf{v}|_p\, \|f\|_\infty$. Besides, $\nabla_{\mathbf{v}} f \in V_p^d$, so it then easily follows that $\|f\|_F = \|f\|_\infty$ for $f \in V_p^d$. Therefore, with $m = 2^s$ and the subspace $X_m = V_p^d \subset \widetilde{F}_p^d$, we obtain $b_m(\widetilde{F}_p^d \hookrightarrow L_\infty) = 1$. Since the sequence of Bernstein numbers is decreasing, we know the first $2^s$ Bernstein numbers.

In order to estimate $\|\nabla_{\mathbf{v}} f\|_\infty$, we first consider partial derivatives. For an index $(k-1)r < i \leq kr$, where $k \in \{1, \ldots, s\}$, for $f \in V_p^d$ and $\mathbf{x} \in [0,1]^d$ we have

$$\begin{aligned}
|\partial_i f(\mathbf{x})| &= \tfrac{1}{r}\, |f(x_1, \ldots, x_{(k-1)r}, 1, \ldots, 1, x_{kr+1}, \ldots, x_d) \\
&\qquad - f(x_1, \ldots, x_{(k-1)r}, 0, \ldots, 0, x_{kr+1}, \ldots, x_d)| \\
&\leq \tfrac{1}{r}\, |f(x_1, \ldots, x_{(k-1)r}, 1, \ldots, 1, x_{kr+1}, \ldots, x_d)| \\
&\qquad + |f(x_1, \ldots, x_{(k-1)r}, 0, \ldots, 0, x_{kr+1}, \ldots, x_d)| \\
&\leq \tfrac{2}{r}\, \|f\|_\infty \\
&\leq d^{-1+1/p}\, \|f\|_\infty .
\end{aligned}$$

For the directional derivative $\nabla_{\mathbf{v}} f$, this gives us

$$\begin{aligned}
\|\nabla_{\mathbf{v}} f\|_\infty &\leq \sum_{i=1}^d |v_i|\, \|\partial_i f\|_\infty \\
&\leq |\mathbf{v}|_1\, d^{-1+1/p}\, \|f\|_\infty \\
&\leq |\mathbf{v}|_p\, \|f\|_\infty .
\end{aligned}$$

□

By Theorem 2.1 (or Theorem 2.16, respectively) we directly obtain the following result on the Monte Carlo complexity.

**Corollary 2.20.** *Consider the approximation problem* $\mathrm{APP} : F_p^d \hookrightarrow L_\infty[0,1]^d$ *with parameter* $1 \leq p < \infty$. *Then the Monte Carlo complexity for achieving an error smaller than* $\varepsilon \leq \frac{1}{30}$ *is bounded from below by*

$$n^{\mathrm{ran,ada}}(\varepsilon, \mathrm{APP}, F_p^d) > 2^{\lfloor \frac{d^{1/p}}{3} \rfloor - 1}\,.$$

*For homogeneous algorithms we have the same complexity for* $\varepsilon = \frac{1}{2}$ *already,*

$$n^{\mathrm{ran,hom}}(\tfrac{1}{2}, \mathrm{APP}, F_p^d) \geq 2^{\lfloor \frac{d^{1/p}}{3} \rfloor - 1}\,.$$

*In the case* $p = 1$, *the problem suffers from the curse of dimensionality. In general, for small* $\varepsilon > 0$, *the* $\varepsilon$*-complexity depends exponentially on* $d^{1/p}$.

Note that the initial error is $e(0, \mathrm{APP}, F_p^d) = 1$, hence the problem is properly normalized. Furthermore, functions from $F_p^d$ can be identified with functions in $F_p^{d+1}$ that are independent from $x_{d+1}$.

The upper bounds actually get close to the lower bound in terms of $d$-dependency. The idea originates from Vybíral [77], where it has been used for slightly different settings, but included a case similar to the case $p = \infty$ here.

**Theorem 2.21.** *For the $L_\infty$-approximation of smooth functions from the classes $F_p^d$, $1 \le p \le \infty$, for $\varepsilon > 0$, we obtain the following upper bounds on the $\varepsilon$-complexity achieved by linear deterministic algorithms,*

$$n^{\mathrm{det,lin}}(\varepsilon, \mathrm{APP}, F_p^d, \Lambda^{\mathrm{all}}) \le \exp\left(\log(d+1)\, \max\left\{\frac{\log \frac{1}{\varepsilon}}{\log 2},\, \mathrm{e}\, d^{1/p}\right\}\right).$$

*In particular, in the case $p > 1$, the problem is $(s,t)$-weakly tractable for all $t > \frac{1}{p}$ and $s > 0$, but not for $t < \frac{1}{p}$. In the case of $p = \infty$, the problem is quasi-polynomially tractable.*

*Proof.* As an algorithm we consider the $k$-th Taylor polynomial, $k \in \mathbb{N}_0$, at the point $\mathbf{x}_0 := (\frac{1}{2}, \ldots, \frac{1}{2})$, then the output function is defined for $\mathbf{x} = \mathbf{x}_0 + \mathbf{v} \in [0,1]^d$ as

$$[A_k(f)](\mathbf{x}) := \sum_{j=0}^{k} \frac{[\nabla_{\mathbf{v}}^j f](\mathbf{x}_0)}{j!} = \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{N}_0^d \\ |\boldsymbol{\alpha}|_1 \le k}} \frac{[D^{\boldsymbol{\alpha}} f](\mathbf{x}_0)}{\boldsymbol{\alpha}!}\, \mathbf{v}^{\boldsymbol{\alpha}}.$$

For this deterministic algorithm we need

$$n(k) := \mathrm{card}(A_k) = \binom{d+k}{d} \le (d+1)^k \tag{2.4.11}$$

partial derivatives of the input $f$ at $\mathbf{x}_0$ as information.[13] The error estimate then is

$$|f(\mathbf{x}) - [A_k(f)](\mathbf{x})| \le \frac{\|\nabla_{\mathbf{v}}^{k+1} f\|_\infty}{(k+1)!},$$

and with $|\mathbf{v}|_p \le \frac{1}{2}\, d^{1/p}$, for the input set $F_p^d$ we obtain the error bound

$$\begin{aligned} e(A_k, F_p^d) &\le \frac{1}{(k+1)!} \left(\frac{d^{1/p}}{2}\right)^{k+1} \\ \text{[Stirling's formula]} \quad &\le \frac{1}{\sqrt{2\,\pi\,(k+1)}} \left(\frac{\mathrm{e}\, d^{1/p}}{2\,(k+1)}\right)^{k+1}. \end{aligned} \tag{2.4.12}$$

[13] Vybíral [77] even shows that the same amount of function values is actually sufficient to approximate the partial derivatives at the point $\mathbf{x}_0$ with arbitrarily high accuracy.
The problem of counting the number of partial derivatives $D^{\boldsymbol{\alpha}} f(\mathbf{x}_0)$ up to the order $|\boldsymbol{\alpha}|_1 \le k$, $\boldsymbol{\alpha} \in \mathbb{N}_0^d$, is equivalent to choosing $d$ numbers $t_1 < \ldots < t_d$ from the set $\{1, \ldots, d+k\}$ by the transformation $\alpha_j := t_j - t_{j-1} - 1$, where $t_0 = 0$. For our purpose, it is sufficient to know that we do not need more than $(d+1)^k$ partial derivatives (like deciding $k$ times in which coordinate direction to derive – or not to derive – the function) since $k$ is very small.

What $k \in \mathbb{N}_0$ should we choose in order to guarantee an error smaller or equal a given tolerance $\varepsilon > 0$? This is ensured for

$$\left(\frac{\mathrm{e}\, d^{1/p}}{2\,(k+1)}\right)^{k+1} \leq \varepsilon \quad \Leftrightarrow \quad k+1 \geq \frac{\mathrm{e}\, d^{1/p}}{2}\, \varepsilon^{-\frac{1}{k+1}}\,.$$

Note that $(1/\varepsilon)^{1/(k+1)} \leq 2$ for $k+1 \geq \log \frac{1}{\varepsilon} / \log 2$, so choosing

$$k = k(\varepsilon) = \left\lfloor \max\left\{ \frac{\log \frac{1}{\varepsilon}}{\log 2},\, \mathrm{e}\, d^{1/p} \right\} \right\rfloor$$

will give us the guarantee we aim for. By this and (2.4.11), we obtain the theorem on the $\varepsilon$-complexity. □

This upper bound is not optimal in terms of the speed of convergence, which is superpolynomial (as it has already been mentioned in Novak and Woźniakowski [56]). However, together with Corollary 2.20, it shows that in these cases randomization does not help to improve the tractability classification of the problems. Here, only narrowing the input set affects tractability.

There are several other publications worth mentioning that study the tractability of the approximation of smooth functions in the worst case setting. Weimar [79] discusses several settings with weighted Banach spaces. Xu [80] considers the $L_p$-approximation for $1 \leq p < \infty$ and the same input set $F^d$ as in Novak and Woźniakowski [56].

# Chapter 3

# Uniform Approximation of Functions from a Hilbert Space

We study the $L_\infty$-approximation of functions from Hilbert spaces with linear functionals $\Lambda^{\mathrm{all}}$ as information. Based on a fundamental Monte Carlo approximation method (originally for finite dimensional input spaces) which goes back to Mathé [46], see Section 3.2.1, we propose a function approximation analogue to standard Monte Carlo integration, now using "Gaussian linear functionals" as random information, see Section 3.3.1. This method is intended to break the curse of dimensionality which holds in the deterministic setting. The analysis relies on the theory of Gaussian fields, see Section 3.3.3, some theory of reproducing kernel Hilbert spaces is needed as well, see Section 3.3.2. Using a known proof technique for lower bounds in the worst case setting, see Section 3.3.4, we can prove the curse of dimensionality for the deterministic approximation of functions from unweighted periodic tensor product Hilbert spaces, whereas for the randomized approximation we can show polynomial tractability under certain assumptions, see Section 3.4.2 for this particular application. A specific example are Korobov spaces, see Theorem 3.19.

## 3.1 Motivation and the General Setting

For the integration problem with standard information $\Lambda^{\mathrm{std}}$, it is known since more than half a century that randomization can speed-up the order of convergence. For example, for $r$-times continuously differentiable functions

$$F_r^d := \{f \in C^r([0,1]^d) \mid \|D^{\boldsymbol{\alpha}} f\|_\infty \leq 1 \text{ for } \boldsymbol{\alpha} \in \mathbb{N}_0^d \text{ with } |\boldsymbol{\alpha}|_1 \leq r\},$$

one can show

$$e^{\mathrm{ran}}(n, \mathrm{INT}, F_r^d) \asymp n^{-r/d-1/2} \prec e^{\mathrm{det}}(n, \mathrm{INT}, F_r^d) \asymp n^{-r/d},$$

where the hidden constants depend on $r$ and $d$, see for instance the lecture notes of Novak [51, Secs 1.3.8/9 and 2.2.9], the original result is due to Bakhvalov 1959 [6]. We see that for fixed smoothness and high dimensions the deterministic rate gets arbitrarily bad, whereas for Monte Carlo methods we have a guaranteed rate of

convergence of $n^{-1/2}$. Even worse, in the deterministic setting upper bounds are achieved by product rules that use $n = m^d$ function values on a regular grid as information, which for high dimensions is of no practical use. The classical approach to lower bounds – when proving the rate of convergence, for both settings – involved constants that are exponentially small in $d$. Within a recent result, Hinrichs, Novak, Ullrich, and Woźniakowski [27] proved the curse of dimensionality for these classes of $C^r$-functions in the deterministic setting,

$$n^{\mathrm{det}}(\varepsilon, \mathrm{INT}, F_r^d) \geq c_r^d \, (d/\varepsilon)^{d/r} \quad \text{for all } d \in \mathbb{N} \text{ and } 0 < \varepsilon < 1/2,$$

moreover, they proved that product rules are really the best what we can do. In contrast to this, the problem is strongly polynomially tractable in the randomized setting by simple means of the classical Monte Carlo method

$$M_n(f) := \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_i) \,, \tag{3.1.1}$$

where the $\mathbf{X}_i$ are iid uniformly distributed on the domain $[0,1]^d$. Here, we have the bound $e(M_n, F_r^d) \leq n^{-1/2}$, and therefore

$$n^{\mathrm{ran}}(\varepsilon, F_r^d) \leq \lceil \varepsilon^{-2} \rceil \quad \text{for } \varepsilon > 0.$$

When aiming for the optimal rate $n^{-r/d-1/2}$, a proof would usually consider algorithms that use exponentially in $d$ many function values, so in many high-dimensional cases the standard Monte Carlo method might be the best approach to a practical solution.

This observation raises the question whether there are approximation problems where randomization significantly reduces the complexity. Even more, can we find a comparably simple Monte Carlo method that breaks the curse of dimensionality? The short answer is: Yes, we can – at least for some problems.

Throughout this chapter we consider linear problems

$$S : \widetilde{F} \hookrightarrow G$$

with the input set $F$ being the unit ball of $\widetilde{F}$, allowing algorithms to use arbitrary continuous linear functionals $\Lambda^{\mathrm{all}}$ for information. The latter is a crucial assumption for the new upper bounds based on a fundamental Monte Carlo approximation method, see Proposition 3.1 in Section 3.2.1 below. Whilst for the introductory part Section 3.2 the input space $\widetilde{F}$ is not necessarily a Hilbert space, this will be the case for Sections 3.3 and 3.4. The examples we present towards the end of the chapter are all with the output space $G = L_\infty$, but it should also be possible to consider embeddings into classical smoothness spaces $C^r$.

## 3.2 Introduction to Randomized Approximation

The most important part of this introduction is Section 3.2.1 with the fundamental Monte Carlo approximation method Proposition 3.1, an idea which is due to

Mathé [46]. Sections 3.2.2, 3.2.3, and 3.2.4, give an overview of different settings where this method can be applied. The reader can decide to skip these sections and go directly to Section 3.3 where we collect tools for the tractability analysis of function approximation in both, the deterministic and the randomized setting. These methods are applied to exemplary problems in Section 3.4.

Still, those last three sections of the present introductory part (Sections 3.2.2–3.2.4) may be helpful to gain some deeper insight into the potential of randomized approximation and the historical background. Within Section 3.2.2 we consider finite dimensional sequence spaces and summarize what is known about this topic. Recovery of sequences is a keystone for the understanding of how the speed of convergence can be enhanced by randomization for function space embeddings, a short overview on that issue is to be found in Section 3.2.3. Finally, in Section 3.2.4 we discuss a sequence space model for $d$-dependent problems where randomization breaks the curse of dimensionality. This example will give us strong indication that we should restrict ourselves to Hilbert spaces $\mathcal{H}^d$ of $d$-variate functions as input spaces, and $L_\infty$ as the target space, in search for examples where randomization helps to break the curse.

### 3.2.1 A Fundamental Monte Carlo Approximation Method

The following result originates from Mathé [46] and is a key component for the Monte Carlo approximation of Hilbert space functions. Here we keep it a little more general than in the original paper, where the output space was a sequence space $\ell_q^m$ with $q > 2$.

**Proposition 3.1.** *Let $S : \ell_2^m \to G$ be a linear operator between normed spaces and consider the unit ball $B_2^m \subset \ell_2^m$ as the input set. Let the information mapping $N^\omega = N$ be a random $(n \times m)$-Matrix with entries $N_{ij} = \frac{1}{\sqrt{n}} X_{ij}$, where the $X_{ij}$ are independent standard Gaussian random variables. Then $A_n^\omega := S\, N^\top N$ defines a linear rank-$n$ Monte Carlo method ($\phi^\omega(\mathbf{y}) = S\, N^\top \mathbf{y}$) and its error is bounded from above by*

$$e(A_n, S : \ell_2^m \to G) \le \frac{2}{\sqrt{n}}\, \mathbb{E}\, \|S\mathbf{X}\|_G$$

*where $\mathbf{X}$ is a standard Gaussian vector in $\mathbb{R}^m$.*

*Proof.* Note that $A_n$ is an unbiased linear Monte Carlo algorithm. To see this, let $\mathbf{x} \in \mathbb{R}^m$, then

$$\mathbb{E}(N^\top N\mathbf{x})(i) = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{m} \underbrace{\mathbb{E}\, X_{ji} X_{jk}}_{=\delta_{ik}}\, x_k = x_i\,,$$

i.e. $\mathbb{E}\, N^\top N\mathbf{x} = \mathbf{x}$, and by linearity of $S$ we have $\mathbb{E}\, A_n^\omega \mathbf{x} = S\mathbf{x}$.

We start from the definition of the error for an input $\mathbf{x} \in \ell_2^m$,

$$e(A_n, \mathbf{x}) = \mathbb{E}\, \|S\mathbf{x} - SN^\top N\mathbf{x}\|_G\,.$$

Now, let $M$ be an independent copy of $N$. We write $\mathbb{E}'$ for expectations with respect to $M$, and $\mathbb{E}$ with respect to $N$. Using $\mathbb{E}' M^\top M = \mathrm{id}_{\mathbb{R}^m}$ and $\mathbb{E}' M = 0$, we can write

$$\begin{aligned} &= \mathbb{E} \,\|\, \mathbb{E}' \, S(M^\top M - M^\top N + N^\top M - N^\top N)\mathbf{x}\|_G \\ &\overset{\Delta\text{-ineq.}}{\leq} 2\,\mathbb{E}\,\mathbb{E}' \left\| S \left( \frac{M+N}{\sqrt{2}} \right)^\top \left( \frac{M-N}{\sqrt{2}} \right) \mathbf{x} \right\|_G . \end{aligned}$$

The distribution of $(M, N)$ is identical to that of $\left(\frac{M+N}{\sqrt{2}}, \frac{M-N}{\sqrt{2}}\right)$, therefore

$$= 2\,\mathbb{E}\,\mathbb{E}' \, \|SN^\top M\mathbf{x}\|_G.$$

Here, $M\mathbf{x}$ is a Gaussian vector distributed like $\frac{\|\mathbf{x}\|_2}{\sqrt{n}}\,\mathbf{Y}$ with $\mathbf{Y}$ being a standard Gaussian vector on $\mathbb{R}^n$. So we continue, $\mathbb{E}'$ now denoting the expectation with respect to $\mathbf{Y}$,

$$= \frac{2\|\mathbf{x}\|_2}{\sqrt{n}}\,\mathbb{E}\,\mathbb{E}' \, \|SN^\top \mathbf{Y}\|_G.$$

For fixed $\mathbf{Y}$, the distribution of $N^\top \mathbf{Y}$ is identical to that of $\frac{\|\mathbf{Y}\|_2}{\sqrt{n}}\,\mathbf{X}$ with $\mathbf{X}$ being a standard Gaussian vector on $\mathbb{R}^m$. Let $\mathbb{E}$ denote the expectation with respect to $\mathbf{X}$. By Fubini's theorem we get

$$= \frac{2\|\mathbf{x}\|_2}{n}\,\mathbb{E}'[\|\mathbf{Y}\|_2\,\mathbb{E}\,\|S\mathbf{X}\|_G] .$$

Using $\mathbb{E}'\,\|\mathbf{Y}\|_2 \leq \sqrt{\mathbb{E}'\,\|\mathbf{Y}\|_2^2} = \sqrt{n}$, we finally obtain

$$e(A_n, \mathbf{x}) \leq \frac{2\|\mathbf{x}\|_2}{\sqrt{n}}\,\mathbb{E}\,\|S\mathbf{X}\|_G .$$

□

**Remark 3.2** (Properties of the fundamental function approximation method)**.** As mentioned within the proof of the error bound, the algorithm is *unbiased*, that is, $\mathbb{E}\,A_n^\omega \mathbf{x} = S\mathbf{x}$ for $\mathbf{x} \in \ell_2^m$.

However, in general the method is *non-interpolatory* since for non-trivial problems $S$ with positive probability the output will be outside the image $S(B_2^m)$ of the input set $B_2^m \subset \ell_2^m$, which is the unit ball. If the solution operator $S$ is injective, then the output is the solution for $N^\top N\mathbf{x} \in \ell_2^m$, and one can show

$$\sqrt{\mathbb{E}\,\|N^\top N\mathbf{x}\|_2^2} = \sqrt{1 + \frac{m+1}{n}}\,\|\mathbf{x}\|_2 .$$

We will put this to an extreme in Section 3.3.1. Applied to function approximation problems, the method will produce an output function for which the Hilbert norm is almost surely infinite. This means that the output does not only lie outside of the input set, but it actually drops out of the input space $\mathcal{H}$, see Remark 3.4 for the general phenomenon, and Remark 3.20 on the loss of smoothness in the particular context of Korobov spaces.

### 3.2.2 Methods for the Recovery of Sequences

We consider again the identity operator between sequence spaces,

$$\mathrm{APP} : \ell_p^M \hookrightarrow \ell_q^M \,,$$

where $1 \le p, q \le \infty$ and $M \in \mathbb{N}$, compare Section 2.4.1 where lower bounds have been discussed. Now, we summarize what is known on upper bounds. In addition to well-known deterministic bounds, we owe linear Monte Carlo results to Mathé 1991 [46], and non-linear Monte Carlo estimates to Heinrich 1992 [22].

What is the basic structure of deterministic and randomized algorithms, depending on the parameters $p$ and $q$? In what cases does randomization help?

#### The Case $1 \le q \le p \le \infty$ – Practically Complete Information Needed

The simplest case is $1 \le q \le p \le \infty$, where the initial error by simple norm estimates is

$$e(0, \ell_p^M \hookrightarrow \ell_q^M) = M^{1/q-1/p} \,.$$

If we allow to use $n$ information functionals, it is optimal to simply compute the first $n$ entries of the input vector and set the other entries to 0 for the approximant, which gives us

$$e(n, \ell_p^M \hookrightarrow \ell_q^M) = (M-n)^{1/q-1/p} \,,$$

see Pietsch [63, Thm 7.2] for a proof. This is a fairly small reduction of the initial error. At best, we gain a factor at most 2 for $n \le \frac{M}{2}$, in case $p = q$ any $n < M$ will be useless. Randomization and adaption does not help a lot, see page 40 for a deeper discussion. So basically, in this case we can rely on deterministic linear methods.

We now discuss several cases for $1 \le p < q \le \infty$, where the initial error is

$$e(0, \ell_p^M \hookrightarrow \ell_q^M) = 1 \,.$$

#### The Case $1 \le p < q \le 2$ – Non-Linear Deterministic Methods

In the case $p = 1$ and $q = 2$, optimal deterministic error bounds are obtained by non-linear methods with a subtly chosen non-adaptive (that is linear) information mapping $N$. For an information $\mathbf{y} = N\mathbf{x} \in \mathbb{R}^n$ obtained for an input $\mathbf{x} \in \ell_1^M$, one then finds an output $\mathbf{z} \in \mathbb{R}^M$ by $\ell_1$-minimization,

$$\phi(\mathbf{y}) := \operatorname*{argmin}_{\substack{\mathbf{z} \in \mathbb{R}^M \\ N\mathbf{z} = \mathbf{y}}} \|\mathbf{z}\|_1 \,.$$

This definition of the output simply guarantees that the algorithm is interpolatory. Indeed, by construction, $\|\mathbf{z}\|_1 \le \|\mathbf{x}\|_1$, and for inputs $\mathbf{x}$ from the unit ball, which is the input set, the output $\mathbf{z}$ will also be from that input set. Furthermore, it gives the same information. The structure of this method reflects that for linear problems in the deterministic setting interpolatory algorithms based on non-adaptive information are optimal up to a factor 2, see the book on IBC by Traub et al. [73, pp. 51–53 and 57–67]

for further details. In this particular case linear algorithms are far worse than interpolatory algorithms.

The crucial point is to find a good information mapping $N$. Several non-constructive ways are known, e.g. taking an $n \times M$-Matrix with independent standard Gaussian entries, then with positive probability it will have the properties that ensure the up to a constant optimal error bounds

$$e(\phi \circ N, \ell_1^M \hookrightarrow \ell_2^M) \leq C \min\left\{1, \sqrt{\frac{1+\log \frac{M}{n}}{n}}\right\}, \tag{3.2.1}$$

where $C > 0$ is a numerical constant, see Foucart and Rauhut [17, Chap 10] for a proof. Almost surely, the matrix $N$ will be such that the $\ell_1$-minimization is solved by a unique $\mathbf{z} \in \mathbb{R}^M$. Computing $\phi(\mathbf{y})$, in fact, is a linear optimization problem, see Foucart and Rauhut [17, Sec 3.1]. Even more generally, for the parameter range $1 = p < q \leq 2$ and $n < M$, the same algorithms give up to a constant optimal error rates

$$e^{\mathrm{det}}(n, \ell_1^M \hookrightarrow \ell_q^M) \asymp \min\left\{1, \left(\frac{1+\log \frac{M}{n}}{n}\right)^{1-1/q}\right\}. \tag{3.2.2}$$

On the other hand, the lower bounds known for the Monte Carlo setting actually state

$$e^{\mathrm{ran}}(n, \ell_1^M \hookrightarrow \ell_q^M) \succeq n^{-(1-1/q)} \quad \text{for } n \leq \tfrac{M}{2},$$

compare Section 2.4.1. As discussed there, it seems odd for the Monte Carlo error to be independent from the size $M$ of the problem, so we conjecture that randomization may not help significantly in this setting. The gap between the lower and the upper bound is logarithmic in $\frac{M}{n}$.

In Foucart and Rauhut [17, p. 327] one can also find a summary on the worst case error for $1 < p < q \leq \infty$, it is based on results from Kashin 1981 [31]. The basic structure of algorithms in that case again is that the information will be non-adaptive and the output interpolatory, which can be achieved by $\ell_p$-minimization (instead of $\ell_1$-minimization). For simplicity, we only cite the error for $1 < p < q = 2$,

$$e^{\mathrm{det}}(n, \ell_p^M \hookrightarrow \ell_2^M) \asymp \min\left\{1, \frac{M^{1-1/p}}{\sqrt{n}}\right\} \quad \text{for } n < M, \tag{3.2.3}$$

where the hidden constants may depend on $p$. This stands in contrast to the best known lower bounds on the Monte Carlo error from Section 2.4.1,

$$e^{\mathrm{ran}}(n, \ell_p^M \hookrightarrow \ell_2^M) \succeq n^{-(1/p-1/2)} \quad \text{for } n \leq \tfrac{M}{2},$$

which exhibit a polynomial gap of a factor $(\frac{M}{n})^{1-1/p} \leq \sqrt{\frac{M}{n}}$. Still, we conjecture that randomization will not help a lot in the parameter range $1 \leq p < q \leq 2$.

**The Case $2 \le p < q \le \infty$ – Linear Monte Carlo Approximation**

In the case $p = 2$ and $q = \infty$, since Smolyak 1965 [72] it is well known that

$$e^{\mathrm{det}}(n, \ell_2^M \hookrightarrow \ell_\infty^M) = \sqrt{\frac{M-n}{M}}\,, \tag{3.2.4}$$

see Example 3.11 for more details and references. The optimal algorithm is an orthogonal rank-$n$ projection. In particular for $n < \frac{M}{2}$, the deterministic error cannot go below $\frac{\sqrt{2}}{2}$. By norm estimates[1] we obtain that this lower bound holds in general for $2 \le p < q \le \infty$,

$$e^{\mathrm{det}}(n, \ell_p^M \hookrightarrow \ell_q^M) \ge \tfrac{\sqrt{2}}{2} \quad \text{for } n \le \tfrac{M}{2}.$$

Since this is no significant reduction of the initial error, practically, for the deterministic setting we have the choice between full information $n = M$, or accepting the initial error.

In this parameter range, however, it is helpful to apply the fundamental linear Monte Carlo method from Mathé [47], see Proposition 3.1, as long as $n$ is big enough for that the method's error does not exceed the initial error 1. The Monte Carlo algorithm $A_n = N^\top N$, where $N$ is an $(n \times M)$-matrix with independent zero-mean Gaussian entries of variance $\frac{1}{n}$, has the error

$$e(A_n, \ell_p^M \hookrightarrow \ell_q^M) \le M^{1/2-1/p}\, e(A_n, \ell_2^M \hookrightarrow \ell_q^M) \le 2\, M^{1/2-1/p}\, \frac{\mathbb{E}\,\|\mathbf{X}\|_q}{\sqrt{n}},$$

where $\mathbf{X}$ is a standard Gaussian vector in $\mathbb{R}^M$. Using the norm estimates for Gaussian vectors, see Lemma A.9, we obtain

$$e^{\mathrm{ran}}(n, \ell_p^M \hookrightarrow \ell_q^M) \preceq \begin{cases} \min\left\{1, M^{1/2-1/p+1/q}/\sqrt{n}\right\} & \text{for } 2 \le p < q < \infty, \\ \min\left\{1, M^{1/2-1/p}\sqrt{\frac{1+\log M}{n}}\right\} & \text{for } 2 \le p < q = \infty. \end{cases} \tag{3.2.5}$$

Here, the hidden constant may depend on $q$. Comparing this to the best known lower bounds, see Section 2.4.1, where for $n \le \frac{M}{2}$ we have

$$e^{\mathrm{ran}}(n, \ell_p^M \hookrightarrow \ell_q^M) \succeq \begin{cases} n^{-(1/p-1/q)} & \text{for } 1 \le p < q < \infty, \\ n^{-1/p}\sqrt{1+\log n} & \text{for } 1 \le p < q = \infty, \end{cases} \tag{3.2.6}$$

we observe a gap which is at least logarithmic in $M$ (for $p = 2$ and $q = \infty$), and can grow up to a factor of almost order $\sqrt{\frac{M}{n}}$ (the limiting case is $p \to \infty = q$). Once more, this gap seems to be a deficiency of the lower bounds and not of the algorithms proposed, especially in the case $p = 2$ and $q = \infty$.

[1] With $2 \le p$, for the input set being the unit ball, we have $B_2^M \subseteq B_p^M$. On the other hand, for $q \le \infty$ the error measuring norms are related by $\|\cdot\|_\infty \le \|\cdot\|_q$, making the problem even more difficult for $q < \infty$.

**The Case $1 \le p < 2 < q \le \infty$ – Non-Linear Monte Carlo Approximation**

It is easier to approximate with the error measured in an $\ell_q$-norm for $2 < q \le \infty$ than with respect to the $\ell_2$-norm, so

$$e^{\det}(n, \ell_p^M \hookrightarrow \ell_q^M) \le e^{\det}(n, \ell_p^M \hookrightarrow \ell_2^M)\,. \tag{3.2.7}$$

In view of the preceding paragraph it is not surprising that for $n \le \frac{M}{2}$ the order of the worst case error cannot be improved. For $1 < p < 2 < q \le \infty$ we have the estimate

$$e^{\det}(n, \ell_p^M \hookrightarrow \ell_q^M) \asymp \min\left\{1, \frac{M^{1-1/p}}{\sqrt{n}}\right\} \quad \text{for } n \le \tfrac{M}{2}, \tag{3.2.8}$$

see Foucart and Rauhut [17, p. 327]. All in all, we can use the same algorithms that we used for $\ell_p^M \hookrightarrow \ell_2^M$. Only for $p = 1$ and $2 < q \le \infty$, best known lower bounds, see Foucart and Rauhut [17, Thm 10.10], do not match the upper bounds we obtain by (3.2.7). In this case, for $n < M$ we have

$$\min\left\{1, \left(\frac{1+\log\frac{M}{n}}{n}\right)^{1-1/q}\right\} \preceq e^{\det}(n, \ell_1^M \hookrightarrow \ell_q^M) \preceq \min\left\{1, \sqrt{\frac{1+\log\frac{M}{n}}{n}}\right\}. \tag{3.2.9}$$

Randomization, however, enables us to exploit the advantage of measuring the error in an $\ell_q$-norm. Namely, we combine the non-linear deterministic algorithms that we have for $\ell_p \hookrightarrow \ell_2$ with the linear Monte Carlo approximation for $\ell_2 \hookrightarrow \ell_q$, this idea is contained in Heinrich [22, Prop 3]. In detail, split the cost $n = n_1 + n_2$, collecting information $N^\omega \mathbf{x} = (N_1\mathbf{x}, N_2^\omega \mathbf{x}) = (\mathbf{y}_1, \mathbf{y}_2) = \mathbf{y}$ for $\mathbf{x} \in \mathbb{R}^M$, where $N_1$ is an $(n_1 \times m)$-matrix as we would choose it for $\ell_p \hookrightarrow \ell_2$, and $N_2^\omega$ is a random $(n_2 \times m)$-matrix with iid zero-mean Gaussian entries of variance $1/n_2$. The output is generated in two steps. In the first step we compute a rough deterministic approximant

$$\mathbf{z}_1 = \phi_1(\mathbf{y}_1) := \operatorname*{argmin}_{\substack{\mathbf{z}\in\mathbb{R}^m \\ N_1\mathbf{z}=\mathbf{y}_1}} \|\mathbf{z}\|_p\,,$$

in the second step we generate the refined output

$$\phi^\omega(\mathbf{y}_1, \mathbf{y}_2) := \mathbf{z}_1 + [N_2^\omega]^\top(\mathbf{y}_2 - N_2^\omega \mathbf{z}_1)\,.$$

The error for $\|\mathbf{x}\|_p \le 1$ can be estimated as

$$\begin{aligned} e((\phi^\omega \circ N^\omega)_{\omega\in\Omega}, \ell_p^M \hookrightarrow \ell_q^M, \mathbf{x}) &= \mathbb{E}\,\|\mathbf{x} - \mathbf{z}_1 - (N_2^\omega)^\top N_2^\omega(\mathbf{x} - \mathbf{z}_1)\|_q \\ &\le e(([N_2^\omega]^\top N_2^\omega))_{\omega\in\Omega}, \ell_2^M \hookrightarrow \ell_q^M)\,\|\mathbf{x} - \mathbf{z}_1\|_2 \\ &\le e(([N_2^\omega]^\top N_2^\omega)_{\omega\in\Omega}, \ell_2^M \hookrightarrow \ell_q^M)\, e(\phi_1 \circ N_1, \ell_p^M \hookrightarrow \ell_2^M)\,. \end{aligned}$$

One could go with $n_1 = n_2$, then using (3.2.1) or (3.2.3), respectively, together

with (3.2.5), we obtain[2]

$$e^{\mathrm{ran}}(n, \ell_p^M \hookrightarrow \ell_q^M) \preceq \frac{1}{n} \begin{cases} \sqrt{(1+\log M)\left(1+\log \frac{M}{n}\right)} & \text{for } 1 = p \text{ and } q = \infty, \\ M^{1-1/p}\sqrt{1+\log M} & \text{for } 1 < p \leq 2 \text{ and } q = \infty, \\ M^{1/q}\sqrt{1+\log \frac{M}{n}} & \text{for } 1 = p \text{ and } 2 < q < \infty, \\ M^{1-1/p+1/q} & \text{for } 1 < p \leq 2 < q < \infty, \end{cases} \tag{3.2.10}$$

where the hidden constant may depend on $q$. However, if $n$ is too small, it might be better to omit the second step and choose $n_1 = n$, $n_2 = 0$, thus simply taking $\phi_1 \circ N_1$ as a deterministic algorithm that achieves the bound from (3.2.8). Also note that for $n \prec M^{2-2/p}$ or $n \prec 1 + \log \frac{M}{n}$, respectively, this estimate is not optimal, and one should rather use the relation between the case $p < 2$ and the case of the input space being $\ell_2^M$,

$$e^{\mathrm{ran}}(n, \ell_p^M \hookrightarrow \ell_q^M) \leq e^{\mathrm{ran}}(n, \ell_2^M \hookrightarrow \ell_q^M) \overset{(3.2.5)}{\preceq} \frac{1}{\sqrt{n}} \begin{cases} M^{1/q} & \text{for } q < \infty, \\ \sqrt{1+\log M} & \text{for } q = \infty. \end{cases} \tag{3.2.11}$$

Again, known lower bounds (3.2.6) do not reflect the size $M$ of the problem.

### 3.2.3 Speeding up the Convergence for Function Approximation

Several examples are known where the order of convergence can be improved by randomization. Heinrich [22] considered Sobolev embeddings

$$\mathrm{APP} : W_p^r([0,1]^d) \hookrightarrow L_q([0,1]^d)\,,$$

where $r, d \in \mathbb{N}$, and $1 \leq p, q \leq \infty$, with the compactness condition $\frac{r}{d} > \frac{1}{p} - \frac{1}{q}$. Here, $W_p^r([0,1]^d)$ denotes the Sobolev space of smoothness $r$, that is the space of all functions $f \in L_p([0,1]^d)$ such that for all $\boldsymbol{\alpha} \in \mathbb{N}_0^d$, with $|\boldsymbol{\alpha}|_1 \leq r$, the partial derivatives $D^{\boldsymbol{\alpha}} f$ exist in a weak sense and belong to $L_p$. We consider the norm[3]

$$\|f\|_{W_p^r} := \begin{cases} \left( \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{N}_0^d \\ |\boldsymbol{\alpha}|_1 \leq r}} \|D^{\boldsymbol{\alpha}} f\|_p^p \right)^{1/p} & \text{for } 1 \leq p < \infty, \\ \max\limits_{\substack{\boldsymbol{\alpha} \in \mathbb{N}_0^d \\ |\boldsymbol{\alpha}|_1 \leq r}} \|D^{\boldsymbol{\alpha}} f\|_\infty & \text{for } p = \infty, \end{cases}$$

see for example Evans [14, Sec 5.2], or Triebel [74, Sec 2.3, esp. Thm 2.3.3 and Rem 2.3.3/5].

[2] Heinrich [22, Cor 3] contains only the case $p = 1$ since the other cases were not needed for the application to Sobolev embeddings.

[3] Concerning the order of convergence, any equivalent norm will give the same results.

For the deterministic setting we refer to Vybíral [75, Thm 4.12]. For simplicity we only cite the result for smoothness $r > d$,

$$e^{\mathrm{det}}(n, W_p^r([0,1]^d) \hookrightarrow L_q([0,1]^d)) \asymp \begin{cases} n^{-r/d} & \text{for } 1 \le q \le p \le \infty, \\ & \text{or } 1 \le p < q \le 2, \\ n^{-r/d+1/2-1/q} & \text{for } 1 \le p < 2 < q \le \infty, \\ n^{-r/d+1/p-1/q} & \text{for } 2 \le p \le q \le \infty. \end{cases}$$

For the one-dimensional case see also Pinkus [64, Chap VII]. From Heinrich [22, Thm 2] we know the randomized setting for smoothness $r > d$,

$$e^{\mathrm{ran}}(n, W_p^r([0,1]^d) \hookrightarrow L_q([0,1]^d)) \preceq \begin{cases} n^{-r/d} & \text{for } 1 \le p \le \infty \\ & \text{and } 1 \le q < \infty, \\ & \text{or } p = q = \infty, \\ n^{-r/d}\sqrt{1+\log n} & \text{for } 1 \le p < q = \infty. \end{cases}$$

Heinrich also proved lower bounds for the adaptive Monte Carlo setting that match the rate of the upper bounds – except for the case $1 \le q < p = \infty$, where a logarithmic gap of a factor $1/\sqrt{1+\log n}$ occurrs.[4] Note that in all cases the hidden constants may depend on $r$, $d$, $p$, and $q$. In comparison of these two results, Heinrich could show that, for $p < q$ and $2 < q \le \infty$, randomized algorithms can improve the rate of convergence by a factor that can reach almost the order $1/\sqrt{n}$, the most prominent case is $p = 2$ and $q = \infty$. This phenomenon was already known to Mathé [46] in the 1-dimensional case.

Similar gaps between the Monte Carlo and the worst case error have been found in Fang and Duan [15] for multi-variate periodic Sobolev spaces with bounded mixed derivative.[5] Again, gaps occur in parameter settings where we also know that randomization can help for the sequence space embedding $\ell_p^M \hookrightarrow \ell_q^M$. This is not surprising since the estimates for function space embeddings heavily rely on results for sequence space embeddings. In order to illustrate the connection to sequence spaces, let us outline the methods of discretization.

For lower bounds one usually finds $m$-dimensional subspaces $X_m \subseteq \widetilde{F}$ with $m \ge 2n$ such that the restriction $S|_{X_m}$ resembles the sequence space embedding $\ell_p^m \hookrightarrow \ell_q^m$.

Upper bounds are based on Maiorov's discretization technique [44], where the solution operator

$$S : \widetilde{F} \to G$$

is split into finite rank operators, so-called *blocks*,

$$S = \sum_{i=1}^{\infty} S_i\,, \quad \mathrm{rk}\, S_i = h_i \in \mathbb{N}\,,$$

[4] This gap can actually be closed in the non-adaptive Monte Carlo setting since the lower bounds of the Sobolev embeddings are based on estimates for the sequence space embedding $\ell_p^M \hookrightarrow \ell_q^M$. For this, in the case $1 \le q < p = \infty$, we have a better lower bound (2.4.4) when restricting to non-adaptive methods.

[5] In Fang and Duan [15], while lower bounds hold for adaptive Monte Carlo methods, upper bounds are obtained with non-adaptive but in some cases *non-linear* methods.

that can be related to sequence space embeddings $\ell_p^{h_i} \hookrightarrow \ell_q^{h_i}$ by estimates

$$e^{\mathrm{ran}}(n, S_i) \le \gamma_i \, e^{\mathrm{ran}}(n, \ell_p^{h_i} \hookrightarrow \ell_q^{h_i})$$

with $\gamma_i > 0$. Now, with $n = n_1 + \ldots + n_k$, $k \in \mathbb{N}$, we have

$$\begin{aligned} e^{\mathrm{ran}}(n, S) &\le \|S_{-k}\|_{F \to G} + \sum_{i=1}^{k} e^{\mathrm{ran}}(n_i, S_i) \\ &\le \|S_{-k}\|_{F \to G} + \sum_{i=1}^{k} \gamma_i \, e^{\mathrm{ran}}(n_i, \ell_p^{h_i} \hookrightarrow \ell_q^{h_i}) \,, \end{aligned}$$

where $S_{-k} := S - (S_1 + \ldots + S_k)$. A common shape of the block operators could be that for a Schauder basis $(\psi_j)_{j \in \mathbb{N}}$ of the input space $\widetilde{F}$, we have disjoint index sets

$$\bigsqcup_{i=1}^{\infty} J_i = \mathbb{N}$$

of cardinality $\# J_i = h_i$, such that for $f = \sum_{j=1}^{\infty} a_j \, \psi_j$ we may write

$$S_i(f) := \sum_{j \in J_i} a_j \, S(\psi_j) \,. \tag{3.2.12}$$

This structure can be found in Fang and Duan [15], however, in the case of Heinrich [22] the discretization is based on another decomposition that is described in the book of König [35].

As mentioned before, the hidden constants for the results on the order of convergence may depend on the problem parameters. Indeed, the upper and the lower bounds may differ largely, even exponentially in $d$. In particular, for Heinrich's result, the upper bounds are obtained with $n$ being exponential in $d$.

We want to point out another drawback of splitting the operator, especially in the randomized setting. Let the input space be a Hilbert space $\widetilde{F} = \mathcal{H}$ with orthonormal basis $(\psi_j)_{j \in \mathbb{N}}$, and split the operator $S$ into block operators $S_i$ of the structure (3.2.12). Now, performing Maiorov's technique, we approximate the first $k$ block operators by known methods, using a part $n_i$ of the total information, respectively, where $n = n_1 + \ldots + n_k$. Assume that for each of the blocks, using $n_i$ pieces of information, the fundamental Monte Carlo approximation method from Proposition 3.1 is the best method we know. Then we obtain

$$e^{\mathrm{ran}}(n, S_1 + \ldots + S_k) \le \sum_{i=1}^{k} e^{\mathrm{ran}}(n_i, S_i) \le \sum_{i=1}^{k} \frac{2 \, \mathbb{E} \left\| \sum_{j \in J_i} X_j \, S(\psi_j) \right\|_G}{\sqrt{n_i}} \,,$$

where the $X_j$ are independent standard Gaussian random variables. However, we could apply the fundamental Monte Carlo approximation method directly to the cluster $S_1 + \ldots + S_k$, and obtain the far better estimate

$$e^{\mathrm{ran}}(n, S_1 + \ldots + S_k) \le \frac{2 \, \mathbb{E} \left\| \sum_{i=1}^{k} \sum_{j \in J_i} X_j \, S(\psi_j) \right\|_G}{\sqrt{n}} \,.$$

(Apply the triangle inequality for comparison to the Maiorov type upper bound.) For this reason, for our analysis on breaking the curse, we will take a more direct approach to the problem, see Section 3.3.1.

Let us add one final remark on the type of information. In this chapter we aim for examples of $d$-dependent problems where randomized approximation using information from arbitrary linear functionals $\Lambda^{\mathrm{all}}$ can break the curse of dimensionality. The examples of enhanced speed of convergence were also based on general information $\Lambda^{\mathrm{all}}$. However, randomization can also help in some cases where only function values $\Lambda^{\mathrm{std}}$ are available to the algorithms. This was shown by Heinrich in a series of papers [23, 24, 25], where he studied the randomized approximation of Sobolev embeddings, and discovered cases of low smoothness where randomization can give a speedup over deterministic methods. It is an interesting task for future research to find examples of function approximation problems based on standard information $\Lambda^{\mathrm{std}}$ where randomization can break the curse of dimensionality, or significantly improve the $d$-dependency of a problem.

### 3.2.4 Breaking the Curse - a Sequence Space Modell

Consider the following example with a rather artificial[6] dimensional parameter $d \in \mathbb{N}$,

$$\mathrm{APP} : \ell_2^{2^d} \hookrightarrow \ell_\infty^{2^d}. \tag{3.2.13}$$

The initial error is 1, hence the problem is properly normalized. By (3.2.4) we have

$$n^{\mathrm{det}}(\varepsilon, \ell_2^{2^d} \hookrightarrow \ell_\infty^{2^d}) \geq 2^{d-1} \quad \text{for } 0 < \varepsilon \leq \tfrac{\sqrt{2}}{2},$$

which clearly is the curse of dimensionality. Now, in the randomized setting, by (3.2.5) we have

$$n^{\mathrm{ran}}(\varepsilon, \ell_2^{2^d} \hookrightarrow \ell_\infty^{2^d}) \leq C\, d\, \varepsilon^{-2} \quad \text{for } \varepsilon > 0, \tag{3.2.14}$$

where $C > 0$ is a numerical constant. This means that the problem is polynomially tractable for Monte Carlo methods.

We want to discuss briefly what happens for the problem

$$\mathrm{APP} : \ell_p^{2^d} \hookrightarrow \ell_q^{2^d}$$

in other parameter settings where Monte Carlo methods are known to improve the error significantly, that is for $1 \leq p < q$ and $2 < q \leq \infty$, see Section 3.2.2.

In what cases do we have the curse of dimensionality for the deterministic setting in the first place?
It turns out that, if $p = 1$, then by (3.2.7) for $2 < q \leq \infty$ we have

$$n^{\mathrm{det}}(\varepsilon, \ell_1^{2^d} \hookrightarrow \ell_q^{2^d}) \preceq d\, \varepsilon^{-2},$$

[6] One could regard $\ell_2^{2^d}$ as an $L_2$-space on a Boolean domain $\{0,1\}^d$ equipped with the counting measure $\#$.

which implies polynomial tractability in the deterministic setting already.[7]
For $2 \leq p < q \leq \infty$, in turn, the problem is more difficult than the problem (3.2.13) and we obviously inherit the curse of dimensionality.
Furthermore, in the case $1 < p < 2 < q \leq \infty$, by (3.2.8) we have the estimate

$$n^{\mathrm{det}}(\varepsilon, \ell_p^{2^d} \hookrightarrow \ell_q^{2^d}) \succeq 2^{(2-2/p)\,d} \quad \text{for } 0 < \varepsilon < \varepsilon_0,$$

so in this case deterministic methods suffer from the curse of dimensionality, too.

Now, what do we know about randomized approximation for $p > 1$?
If the target space is altered compared to the original example (3.2.13), i.e. $q < \infty$, or if the input set[8] is extended, that is the case for $p > 2$, then by (3.2.5), or by (3.2.10) and (3.2.11), respectively, we only know the upper bounds for $0 < \varepsilon < \varepsilon_q$,[9]

$$n^{\mathrm{ran}}(\varepsilon, \ell_p^{2^d} \hookrightarrow \ell_q^{2^d}) \preceq \begin{cases} 2^{(1-2/p+2/q)\,d}\,\varepsilon^{-2} & \text{for } 2 \leq p < q < \infty, \\ d\,2^{(1-2/p)\,d}\,\varepsilon^{-2} & \text{for } 2 < p < q = \infty, \\ \min\left\{2^{(1-1/p+1/q)\,d}\,\varepsilon^{-1}, 2^{2d/q}\,\varepsilon^{-2}\right\} & \text{for } 1 < p < 2 < q < \infty, \end{cases}$$

which is still exponential in $d$. We do not know whether the curse of dimensionality actually holds in the randomized setting because the lower bounds we obtain from (3.2.6) will be independent from $d$.
The case left over is $1 < p < 2$ and $q = \infty$. Here, we can break the curse of dimensionality similarly to the original example (3.2.13). Indeed, the $\ell_p$-ball is contained in the $\ell_2$ ball, hence the problem is easier.

To summarize, the $\ell_\infty$ approximation of finite $\ell_p$-sequences with $1 < p \leq 2$ is the case where we know that randomization can break the curse of dimensionality. The most prominent case is $p = 2$ where best known Monte Carlo methods are linear, yet linear methods would suffice to break the curse for $1 < p < 2$ as well.

This sequence space example motivates the restriction to the $L_\infty$-approximation of Hilbert space functions in search of function approximation problems where the curse of dimensionality holds in the worst case setting but polynomial tractability can be found in the Monte Carlo setting.

## 3.3 Tools for Function Approximation

In Section 3.3.1 we put the fundamental Monte Carlo method from Proposition 3.1 to an extreme and obtain a function approximation analogue to standard Monte Carlo integration (3.1.1). Here, we restrict the input set to functions from Hilbert spaces. Lemma 3.3 stated below is still quite general, its specification to $L_\infty$-approximation of functions is the starting point for the study of Gaussian random fields and their

[7] By (3.2.2) we have polynomial tractability for $p = 1 < q \leq 2$ as well, yet with a worse $\varepsilon$-dependency $\varepsilon^{-q/(q-1)}$.

[8] Recall that the input set is the unit ball $B_p^{2^d}$ of $\ell_p^{2^d}$.

[9] The constant $\varepsilon_q > 0$ is actually the hidden constant from the respective error estimates (3.2.5), (3.2.10) and (3.2.11), the hidden constant for the complexity estimates is then $\varepsilon_q^2$, or $\varepsilon_q$ for $1 < p < 2 < q < \infty$.

expected maximum. In preparation for this, in Section 3.3.2 we sketch major elements of the theory of *reproducing kernel Hilbert spaces* (RKHS). Section 3.3.3 then outlines the well established theory of Gaussian fields associated to a RKHS, in particular the technique of majorizing measures due to Fernique, and Dudley's entropy-based estimates. The theory of RKHSs is also useful for the analysis of the worst case setting, for which we need lower bounds in order to show the superiority of Monte Carlo approximation. Section 3.3.4 addresses a general approach to deterministic $L_\infty$-approximation of Hilbert space functions, that approach has been taken by several authors before [12, 40, 61].

### 3.3.1 A Plain Monte Carlo Upper Bound

**Lemma 3.3.** *Consider the linear problem with a compact solution operator*

$$S : \mathcal{H} \to G$$

*from a separable Hilbert space $\mathcal{H}$ into a Banach space $G$, the input set $F \subset \mathcal{H}$ being the unit ball. Assume that we have an orthonormal basis $(\psi_j)_{j\in\mathbb{N}}$ for $\mathcal{H}$ such that the sum $\sum_{j=1}^\infty X_j\, S(\psi_j)$, with independent standard Gaussian random variables $X_j$, converges almost surely in $G$. Then for $n \in \mathbb{N}$ we have*

$$e^{\mathrm{ran}}(n, S, \Lambda^{\mathrm{all}}) \le \frac{2}{\sqrt{n}}\, \mathbb{E}\left\| \sum_{j=1}^\infty X_j\, S(\psi_j) \right\|_G ,$$

*or equivalently, for $\varepsilon > 0$,*

$$n^{\mathrm{ran}}(\varepsilon, S, \Lambda^{\mathrm{all}}) \le \left\lceil 4 \left( \frac{\mathbb{E}\left\| \sum_{j=1}^\infty X_j\, S(\psi_j) \right\|_G}{\varepsilon} \right)^2 \right\rceil .$$

*Proof.* For $m \in \mathbb{N}$ we define the linear Monte Carlo method $A_{n,m} = (A^\omega_{n,m})_\omega$ which, for an input $f \in \mathcal{H}$, returns the output

$$g = A^\omega_{n,m}(f) := \frac{1}{n} \sum_{i=1}^n L^\omega_{i,m}(f)\, g^\omega_{i,m} ,$$

based on the information

$$y_i = L^\omega_{i,m}(f) = \sum_{j=1}^m X_{ij}\, \langle \psi_j, f \rangle_{\mathcal{H}} ,$$

and with elements from the output space

$$g^\omega_{i,m} := \sum_{j=1}^m X_{ij}\, S(\psi_j) .$$

Here, the $X_{ij}$ are independent standard Gaussian random variables. This algorithm is actually the fundamental Monte Carlo method from Proposition 3.1 when restricting $S$

to the subspace $\mathcal{H}_m := \operatorname{span}\{\psi_1, \dots, \psi_m\}$ that can be identified with $\ell_2^m$. Let $P_m$ denote the orthogonal projection onto $\mathcal{H}_m$. Since $S$ is compact, we have

$$\|S\,(\mathrm{id}_{\mathcal{H}} - P_m)\|_{\mathcal{H}\to G} \xrightarrow[m\to\infty]{} 0\,.$$

Then, for elements $f$ from the input set, $\|f\|_{\mathcal{H}} \le 1$, we have

$$\begin{aligned} e((A^\omega_{n,m})_\omega, f) &\le \|S\,(\mathrm{id}_{\mathcal{H}} - P_m)\,f\|_G + e((A^\omega_{n,m})_\omega, P_m f) \\ &\le \|S(\mathrm{id}_{\mathcal{H}} - P_m)\|_{\mathcal{H}\to G} + \frac{2}{\sqrt{n}}\,\mathbb{E}\left\|\sum_{j=1}^{m} X_j\,S(\psi_j)\right\|_G \\ &\xrightarrow[m\to\infty]{} \frac{2}{\sqrt{n}}\,\mathbb{E}\left\|\sum_{j=1}^{\infty} X_j\,S(\psi_j)\right\|_G . \end{aligned}$$

Note that $\mathbb{E}\left\|\sum_{j=1}^{m} X_j\,S(\psi_j)\right\|_G$ is monotonously increasing for $m \to \infty$ since the $X_j$ are independent, see Lemma A.1. □

**Remark 3.4** (Stochastically bounded information and algorithms)**.** With the above lemma it seems natural to consider the idealized method $A_n = A_{n,\infty}$,

$$A_n(f) := \frac{1}{n}\sum_{i=1}^{n} L_i^\omega(f)\,g_i^\omega\,,$$

with information

$$y_i = L_i^\omega(f) := \sum_{j=1}^{\infty} X_{ij}\,\langle \psi_j, f\rangle_{\mathcal{H}}\,,$$

and elements from the output space

$$g_i^\omega := \sum_{j=1}^{\infty} X_{ij}\,S(\psi_j)\,.$$

Observe the similarities with standard Monte Carlo integration (3.1.1). Observe also the important difference that the present approximation method depends on the particular norm of the input space, whereas standard Monte Carlo integration is defined independently from the input set.

Note that, almost surely, $L_i^\omega$ is an unbounded functional, so $L_i^\omega \notin \Lambda^{\mathrm{all}}$. To see this, for fixed $\omega$, consider the sequence $(f_{ik})_{k=1}^\infty$ of normalized Hilbert space elements

$$f_{ik} := \frac{1}{\sqrt{\sum_{j=1}^{k} X_{ij}^2}}\sum_{j=1}^{k} X_{ij}\,S(\psi_j) \in \mathcal{H}\,,$$

where

$$L_i^\omega(f_{ik}) = \sqrt{\sum_{j=1}^{k} X_{ij}^2} \xrightarrow[k\to\infty]{\text{a.s.}} \infty\,.$$

Indeed, from Lemma A.9 we have $\mathbb{E}\, L_i^\omega(f_{ik}) \geq \sqrt{2k/\pi}$. The deviation result Proposition A.5 can be applied, similarly to Corollary A.6, to bound the probability $\mathbb{P}(L_i^\omega(f_{ik}) < a)$ for $a > 0$. The Borel-Cantelli lemma implies that the monotone sequence $(L_i^\omega(f_{ik}))_{k=1}^\infty$ almost surely exceeds any $a > 0$ for sufficiently large $k$.

Specifically for embedding problems $S = \mathrm{APP} : \mathcal{H} \hookrightarrow G$, the functions $g_i^\omega$, and therefore the output as well, are functions defined on the same domain as the input functions, but they are *not* from the original Hilbert space $\mathcal{H}$ (for the same reasons that caused the functionals $L_i^\omega$ to be discontinuous). This underlines the non-interpolatory nature of the fundamental Monte Carlo approximation method, compare Remark 3.2. Yet the functions $g_i^\omega$ correspond to the information functionals $L_i^\omega$, similar to Hilbert space elements representing continuous linear functionals according to the Riesz representation theorem.

Although the functionals $L_i^\omega$ are almost surely discontinuous, for any fixed input $f \in \mathcal{H}$ the random information $L_i^\omega(f)$ is a standard Gaussian random variable with variance $\|f\|_{\mathcal{H}}^2$, hence almost surely finite. Since, by assumption, the $g_i^\omega$ are almost surely defined, we have a method that is almost surely defined for any fixed $f$. Even more, for any fixed $f \in \mathcal{H}$, the idealized algorithm $A_n$ can be approximated with almost sure convergence,

$$A_{n,m}^\omega(f) \xrightarrow[m\to\infty]{\text{a.s.}} A_n^\omega(f)\,.$$

These considerations motivate to extend the class of admissible information functionals $\Lambda^{\mathrm{all}}$ to some class of "stochastically bounded" functionals $\Lambda^{\mathrm{stoch}}$. Actually, this kind of stochastically defined functionals is quite common. For example, the problem of integrating $L_p$-functions by function values is only solvable in the randomized setting since in that case standard information $\Lambda^{\mathrm{std}}$ is discontinuous. Compare also the example from Heinrich and Milla [26] which has been discussed in Section 1.4.

### 3.3.2 Reproducing Kernel Hilbert Spaces

We summarize several facts about *reproducing kernel Hilbert spaces* (RKHS) that are necessary for the numerical analysis of approximation problems

$$\mathrm{APP} : \mathcal{H} \hookrightarrow L_\infty(D)\,,$$

with $\mathcal{H}$ being a separable Hilbert space of functions defined on a domain $D \subset \mathbb{R}^d$. For a general introduction to reproducing kernels, refer to Aronszajn [5]. For an introduction with focus on the associated Gaussian field, see Adler [1, Sec III.2]. The theory of RKHSs is a powerful concept for the analysis of many other numerical settings, e.g. for certain average case problems (see for instance Ritter [68, Chap III]), or when standard information $\Lambda^{\mathrm{std}}$ is considered (see Novak and Woźniakowski [57, 58] for a bunch of examples), it also proves useful for statistical problems (see Wahba [78]).

#### Definition of Reproducing Kernels

We assume function evaluations to be continuous on $\mathcal{H}$. Then by the Riesz representation theorem, for each $\mathbf{x} \in D$ there exists a unique function $K_{\mathbf{x}}(\cdot) \in \mathcal{H}$ such that

for $f \in \mathcal{H}$ we have

$$f(\mathbf{x}) = \langle K_{\mathbf{x}}, f \rangle_{\mathcal{H}} .$$

For $\mathbf{x}, \mathbf{y} \in D$ we define a *symmetric* function

$$K(\mathbf{x}, \mathbf{y}) := K_{\mathbf{y}}(\mathbf{x}) = \langle K_{\mathbf{x}}, K_{\mathbf{y}} \rangle_{\mathcal{H}} = \langle K_{\mathbf{y}}, K_{\mathbf{x}} \rangle_{\mathcal{H}} = K_{\mathbf{x}}(\mathbf{y}) = K(\mathbf{y}, \mathbf{x}) .$$

This function is called the *reproducing kernel* of $\mathcal{H}$. The reproducing kernel $K$ is *positive-semidefinite*, that is, for $\mathbf{x}_1, \ldots, \mathbf{x}_m \in D$ and $a_1, \ldots, a_m \in \mathbb{R}$ we have

$$\sum_{i,j=1}^{m} a_i \, a_j \, K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \sum_{i=1}^{m} a_i \, K_{\mathbf{x}_i}, \sum_{i=1}^{m} a_i \, K_{\mathbf{x}_i} \right\rangle_{\mathcal{H}} \geq 0 .$$

Reversely, any symmetric and positive-semidefinite function $K : D \times D \to \mathbb{R}$ defines an inner product on the linear space $\operatorname{span}\{K_{\mathbf{x}} := K(\mathbf{x}, \cdot) : D \to \mathbb{R} \mid \mathbf{x} \in D\}$ by

$$\left\langle \sum_{i=1}^{m} a_i \, K_{\mathbf{x}_i}, \sum_{j=1}^{n} b_j \, K_{\mathbf{z}_j} \right\rangle_{K} := \sum_{i=1}^{m} \sum_{j=1}^{n} a_i \, b_j \, K(\mathbf{x}_i, \mathbf{z}_j) ,$$

for points $\mathbf{x}_i, \mathbf{z}_j \in D$ and $a_i, b_j \in \mathbb{R}$. Its completion with respect to the corresponding norm uniquely defines a Hilbert space $\mathcal{H}(K)$ which is then called the *reproducing kernel Hilbert space* with kernel $K$. This is the space $\mathcal{H}$ we started with.

### Comparison to the sup-Norm

Knowing the kernel, it is easy to estimate the sup-norm of normalized functions $f \in \mathcal{H}$. Indeed, with $\|f\|_{\mathcal{H}} = 1$ we get

$$\begin{aligned} \|f\|_{\sup} = \sup_{\mathbf{x} \in D} |f(\mathbf{x})| &= \sup_{\mathbf{x} \in D} \langle K_{\mathbf{x}}, f \rangle_{\mathcal{H}} \\ &\leq \sup_{\mathbf{x} \in D} \|K_{\mathbf{x}}\|_{\mathcal{H}} = \sup_{\mathbf{x} \in D} \sqrt{\langle K_{\mathbf{x}}, K_{\mathbf{x}} \rangle_{\mathcal{H}}} = \sup_{\mathbf{x} \in D} \sqrt{K(\mathbf{x}, \mathbf{x})} . \end{aligned} \tag{3.3.1}$$

This is the initial error for the sup-norm approximation. By the Cauchy-Schwarz inequality, the sup-norm of $K$ is determined by values $K(\mathbf{x}, \mathbf{x})$ on the diagonal of $D \times D$,

$$\sup_{\mathbf{x}, \mathbf{z} \in D} |K(\mathbf{x}, \mathbf{z})| \leq \sup_{\mathbf{x} \in D} K(\mathbf{x}, \mathbf{x}) .$$

Therefore from now on we assume $K$ to be bounded.

### Decomposition of Reproducing Kernels and a Worst Case Upper Bound

Let $(\psi_i)_{i \in \mathbb{N}}$ be an orthonormal basis of $\mathcal{H}(K)$, then we can write

$$K(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{\infty} \psi_i(\mathbf{x}) \, \psi_i(\mathbf{z}) . \tag{3.3.2}$$

That way it is easy to determine the reproducing kernel of a subspace $\mathcal{H}' \subset \mathcal{H}$ spanned by $\{\psi_i\}_{i=n+1}^{\infty}$, it is

$$K'(\mathbf{x}, \mathbf{z}) := \sum_{i=n+1}^{\infty} \psi_i(\mathbf{x})\, \psi_i(\mathbf{z}) = K(\mathbf{x}, \mathbf{z}) - \sum_{i=1}^{n} \psi_i(\mathbf{x})\, \psi_i(\mathbf{z})\,. \tag{3.3.3}$$

Therefore, via the linear algorithm

$$A_n(f) := \sum_{j=1}^{n} \langle \psi_j, f \rangle_{\mathcal{H}}\, \psi_j\,,$$

analogously to the initial error (3.3.1), we can estimate the worst case error,

$$e^{\mathrm{det,lin}}(n, \mathcal{H}(K) \hookrightarrow L_\infty(D), \Lambda^{\mathrm{all}}) \le \sup_{\mathbf{x} \in D} \sqrt{K(\mathbf{x}, \mathbf{x}) - \sum_{i=1}^{n} \psi_i(\mathbf{x})\, \psi_i(\mathbf{z})}\,. \tag{3.3.4}$$

Actually, the optimal error can be achieved that way, see Section 3.3.4 for optimality and lower bounds.

### Tensor Product of Reproducing Kernel Hilbert Spaces

Tensor products are a common way in IBC to define multivariate problems, compare for instance Novak and Woźniakowski [55, Sec 5.2], or Ritter [68, Sec VI.2], find many more examples in Novak and Woźniakowski [57, 58].

Let $\mathcal{H}(K_1)$ and $\mathcal{H}(K_2)$ be reproducing kernel Hilbert spaces defined on $D_1$ and $D_2$, respectively. Let $(\varphi_i)_{i \in \mathbb{N}}$ and $(\psi_j)_{j \in \mathbb{N}}$ be corresponding orthonormal bases. Then the tensor product space $\mathcal{H}(K_1) \otimes \mathcal{H}(K_2)$ is the Hilbert space with orthonormal basis $(\varphi_i \otimes \psi_j)_{i,j \in \mathbb{N}}$. Here, $f_1 \otimes f_2$ denotes the tensor product of functions $f_1 \in \mathcal{H}(K_1)$ and $f_2 \in \mathcal{H}(K_2)$,

$$[f_1 \otimes f_2](\mathbf{x}_1, \mathbf{x}_2) := f_1(\mathbf{x}_1)\, f_2(\mathbf{x}_2)\,, \quad \text{defined for } (\mathbf{x}_1, \mathbf{x}_2) \in D_1 \times D_2.$$

With another tensor product function $g_1 \otimes g_2$ of this sort, one easily obtains

$$\langle f_1 \otimes f_2, g_1 \otimes g_2 \rangle_{\mathcal{H}1 \otimes \mathcal{H}_2} = \langle f_1, g_1 \rangle_{\mathcal{H}_1} \langle f_2, g_2 \rangle_{\mathcal{H}_2}\,.$$

Using the representation (3.3.2) for the reproducing kernel, it is easy to see that the reproducing kernel $K$ of the new space is the tensor product of the kernels of the original spaces,

$$K((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{z}_1, \mathbf{z}_2)) := K_1(\mathbf{x}_1, \mathbf{z}_1)\, K_2(\mathbf{x}_2, \mathbf{z}_2)\,,$$

where $(\mathbf{x}_1, \mathbf{x}_2), (\mathbf{z}_1, \mathbf{z}_2) \in D_1 \times D_2$.

### Canonical Metric and Continuity

We consider the *canonical metric*[10] $d_K : D \times D \to [0, \infty)$,

$$d_K(\mathbf{x}, \mathbf{z}) := \|K_\mathbf{x} - K_\mathbf{z}\|_\mathcal{H} = \sqrt{K(\mathbf{x}, \mathbf{x}) - 2\,K(\mathbf{x}, \mathbf{z}) + K(\mathbf{z}, \mathbf{z})}\,.$$

Functions $f \in \mathcal{H}$ are Lipschitz continuous with Lipschitz constant $\|f\|_\mathcal{H}$ with respect to the canonical metric,

$$|f(\mathbf{x}) - f(\mathbf{z})| = |\langle K_\mathbf{x} - K_\mathbf{z}, f\rangle_\mathcal{H}| \le \|K_\mathbf{x} - K_\mathbf{z}\|_\mathcal{H}\,\|f\|_\mathcal{H} = \|f\|_\mathcal{H}\,d_K(\mathbf{x}, \mathbf{z})\,.$$

Hence functions from $\mathcal{H}$ are continuous with respect to any metric $\delta$ on $D$ that is topologically equivalent to the canonical metric $d_K$.

Since we assume $K$ to be bounded, the domain $D$ is bounded with respect to $d_K$,

$$\operatorname{diam}(D) = \sup_{\mathbf{x},\mathbf{z}\in D} d_K(\mathbf{x}, \mathbf{z}) \le 2 \sup_{\mathbf{x}\in D} \sqrt{K(\mathbf{x}, \mathbf{x})}\,.$$

Towards the end of the next section on the boundedness (and continuity) of associated Gaussian fields, we will need the stronger assumption that $D$ is *totally bounded* with respect to $d_K$. That is, for any $r > 0$ the set $D$ can be covered by finitely many balls with radius $r$. In particular, if $D$ is complete with respect to $d_K$, this implies compactness of $D$. Recall that compactness of a set in a metric space implies total boundedness.

## 3.3.3 Expected Maximum of Zero-Mean Gaussian Fields

We discuss zero-mean Gaussian fields and their connection to reproducing kernel Hilbert spaces. All results presented here can be found in the notes by Adler [1]. For some results, Lifshits [43] or Ledoux and Talagrand [41] will also be good references.

**Definition 3.5.** The Gaussian field associated with a reproducing kernel Hilbert space $\mathcal{H}(K)$ is a random function $\Psi : D \to \mathbb{R} \cup \{\pm\infty\}$ such that, for any finite collection of points $\mathbf{x}_1, \ldots, \mathbf{x}_m \in D$, the vector $(\Psi_{\mathbf{x}_i})_{i=1}^m$ is distributed according to a zero-mean Gaussian distribution in $\mathbb{R}^m$, and

$$\operatorname{Cov}(\Psi_\mathbf{x}, \Psi_\mathbf{z}) = K(\mathbf{x}, \mathbf{z})\,.$$

### Series Representation

Let $(\psi_i)_{i=1}^M$ be an orthonormal basis of $\mathcal{H}(K)$, $M \in \mathbb{N} \cup \{\infty\}$.[11] Then the pointwise definition

$$\Psi_\mathbf{x} := \sum_{i=1}^{M} X_i\,\psi_i(\mathbf{x})\,, \tag{3.3.5}$$

[10] If $d_K(\mathbf{x}, \mathbf{z}) = 0$ for some distinct $\mathbf{x} \neq \mathbf{z}$, we only have a semimetric. Then we still obtain a metric for the set of equivalence classes of points that are at distance 0.

[11] The interesting case of course is $M = \infty$. For $M < \infty$ almost sure boundedness is obvious, still, good upper bounds for the expected maximum are of interest.

with $X_i$ being iid standard Gaussian random variables, produces a version of the Gaussian field associated with $\mathcal{H}(K)$. Indeed, the covariance function of $\Psi$ defined that way is the kernel $K$,

$$\mathrm{Cov}(\Psi_{\mathbf{x}}, \Psi_{\mathbf{z}}) = \sum_{i,j=1}^{\infty} (\mathbb{E}\, X_i\, X_j)\, \psi_i(\mathbf{x})\, \psi_j(\mathbf{z}) = \sum_{i=1}^{\infty} \psi_i(\mathbf{x})\, \psi_i(\mathbf{z}) = K(\mathbf{x}, \mathbf{z})\,.$$

### Continuity

Note that for the *canonical metric* $d_K$ of the reproducing kernel Hilbert space $\mathcal{H}(K)$ we have the alternative representation

$$d_K(\mathbf{x}, \mathbf{z}) = \sqrt{\mathbb{E}(\Psi_{\mathbf{x}} - \Psi_{\mathbf{z}})^2}\,.$$

The question of the boundedness of $\Psi$ is closely related to continuity with respect to the canonical metric $d_K$. We say that the Gaussian field with covariance function $K$ is *continuous*, if there exists a version of $\Psi$ with almost surely continuous sample paths.

The classical approach for continuity starts with a countable dense subset $T \subseteq D$ (we therefore assume $D$ to be separable with respect to $d_K$):
If $\Psi$ is continuous on $T$ with respect to the canonical metric $d_K$, it can be uniquely extended to a continuous function on $D$ by the limit

$$\Psi_{\mathbf{x}} := \lim_{T \ni \mathbf{z} \to \mathbf{x}} \Psi_{\mathbf{z}}\,, \tag{3.3.6}$$

see Lifshits [43, Sec 15]. Working with a countable subset $T$ enables us to determine the probability of $\Psi$ being continuous on $T$. This probability is either 0 or 1, see e.g. Adler [1, Thm 3.12]. If $\Psi$ is continuous on $T$ with probability 1, indeed, (3.3.6) defines an almost surely continuous version of the Gaussian field associated with $\mathcal{H}(K)$ according to Definition 3.5.

If the Gaussian field associated with $\mathcal{H}(K)$ has continuous sample paths and the domain $D$ is totally bounded, one can show that the series representation (3.3.5) converges uniformly[12] on $D$ with probability 1, see Adler [1, Thm 3.8]. In the sequel, when talking about $\Psi$, we always mean the series representation, for which uniform convergence implies continuity with respect to $d_K$.

### Boundedness

From now on we assume that the domain $D$ is totally bounded.[13] In this case, for Gaussian fields, continuity with respect to the canonical metric $d_K$ is equivalent to boundedness, see Adler [1, Thm 4.16].

Let $B_K(\mathbf{x}, r)$ denote the closed $d_K$-ball around $\mathbf{x} \in D$ with radius $r > 0$.

The following result can be found in Adler [1, Thm 4.1], it is originally due to Fernique 1975 [16].

[12] That is, the series converges in the $L_\infty$-norm.

[13] If $D$ is not totally bounded, the estimate in Proposition 3.7 (Dudley) will be infinite. Actually, total boundedness of $D$ is a necessary condition for boundedness of the process $\Psi$.

**Proposition 3.6** (Fernique). *Let $\mu$ be any probability measure on $D$, then*

$$\mathbb{E} \sup_{\mathbf{x}\in D} \Psi_{\mathbf{x}} \leq C_{\mathit{Fernique}} \sup_{\mathbf{x}\in D} \int_0^\infty \sqrt{\log(1/\mu(B_K(\mathbf{x}, r)))}\, \mathrm{d}r\,,$$

*where $C_{\mathit{Fernique}} > 0$ is a universal constant.*

From Adler [1, Sec IV.2] one can extract a value $C_{\text{Fernique}} \leq 4\sqrt{3}\,(1/\sqrt{\pi \log 2} + 16)$ $= 115.5462....$ This constant is not optimal. From the book of Ledoux and Talagrand [41, Prop 11.12] (via the Young function $\psi(x) = \exp(x^2) - 1$) we gain the much better estimate $C_{\text{Fernique}} \leq 8\,(2 + 1/\sqrt{2}) = 21.6568....$

A measure for which the right hand side of the above proposition is finite, is called a *majorizing measure* for the metric space $(D, d_K)$. Majorizing measures must be – in a certain way – "well spread" because the integral will be infinite if $\mu(B_K(\mathbf{x}, r)) = 0$ for some $\mathbf{x} \in D$ and $r > 0$. Yet it may be discrete, see the construction for the proof of Proposition 3.7 below. Furthermore, the integral vanishes for $r$ exceeding the diameter of $D$ with respect to $d_K$.

Sometimes it is inconvenient to work with majorizing measures. An alternative way of estimating the maximum of a Gaussian field is based on *metric entropy*. For $r > 0$, let $N(r) = N(r, D, d_K)$ denote the minimal number of $d_K$-balls with radius $r$ needed to cover $D$. The function $H(r) := \log N(r)$ is called the *(metric) entropy* of $D$. The following inequality is based on this quantity, it goes back to Dudley 1973 [13, Thm 2.1].

**Proposition 3.7** (Dudley). *There exists a universal constant $C_{\mathit{Dudley}} > 0$ such that*

$$\mathbb{E} \sup_{\mathbf{x}\in D} \Psi_{\mathbf{x}} \leq C_{\mathit{Dudley}} \int_0^\infty \sqrt{\log N(r)}\, \mathrm{d}r\,.$$

For a direct proof with explicit numerical bound $C_{\text{Dudley}} \leq 4\sqrt{2}$, see Lifshits [43, Sec 14, Thm 1]. In the book of Adler [1, Cor 4.15] Dudley's inequality was derived from Fernique's estimate.

*Idea of a derivation from Fernique's estimate.* By scaling, without loss of generality, the diameter of $D$ is 1. For $k \in \mathbb{N}_0$, let $\{\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{k,N(2^{-k})}\} \subseteq D$ be a minimal collection of points such that $D$ is covered by balls of radius $r = 2^{-k}$,

$$D = \bigcup_{j=1}^{N(2^{-k})} B_K(\mathbf{x}_{k,j}, 2^{-k})\,.$$

Then, defining

$$\mu(E) := \frac{1}{2} \sum_{k=0}^{\infty} 2^{-k} \left[ \frac{1}{N(2^{-k})} \sum_{j=1}^{N(2^{-k})} \mathbb{1}[\mathbf{x}_{k,j} \in E] \right]$$

for $E \subseteq D$, we obtain a majorizing measure and may apply Proposition 3.6 (Fernique), see Adler [1, Lem 4.14] for more details.[14] □

[14] In Adler [1, Lem 4.14] the construction of the measure is less explicit. Check the proof with the construction given here.

Since we are interested in the expected sup-norm of $\Psi$, we also need the following easy lemma, compare Adler [1, Lem 3.1].

**Lemma 3.8.** *For the Gaussian field $\Psi$ with covariance function $K$, we have*

$$\mathbb{E}\,\|\Psi\|_\infty \le \sqrt{\frac{2}{\pi}}\,\inf_{\mathbf{x}\in D}\sqrt{K(\mathbf{x},\mathbf{x})} + 2\,\mathbb{E}\sup_{\mathbf{x}\in D}\Psi_{\mathbf{x}}\,.$$

*Proof.* With $\mathbf{x}_0 \in D$, by the triangle inequality we obtain

$$\mathbb{E}\,\|\Psi\|_\infty \le \mathbb{E}\,|\Psi_{\mathbf{x}_0}| + \mathbb{E}\,\|\Psi - \Psi_{\mathbf{x}_0}\|_\infty\,.$$

Since $\Psi_{\mathbf{x}_0}$ is a zero-mean Gaussian random variable with variance $K(\mathbf{x}_0,\mathbf{x}_0)$, we get

$$\mathbb{E}\,|\Psi_{\mathbf{x}_0}| = \sqrt{\frac{2}{\pi}}\,\sqrt{K(\mathbf{x}_0,\mathbf{x}_0)}\,.$$

For the random field $\Phi_{\mathbf{x}} := \Psi_{\mathbf{x}} - \Psi_{\mathbf{x}_0}$, we have $\Phi_{\mathbf{x}_0} = 0$, so by symmetry

$$\mathbb{E}\,\|\Phi_{\mathbf{x}}\|_\infty = \mathbb{E}\max\{\sup_{\mathbf{x}\in D}\Phi_{\mathbf{x}}, \sup_{\mathbf{x}\in D}(-\Phi_{\mathbf{x}})\} \le \mathbb{E}\sup_{\mathbf{x}\in D}\Phi_{\mathbf{x}} + \mathbb{E}\sup_{\mathbf{x}\in D}(-\Phi_{\mathbf{x}}) = 2\,\mathbb{E}\sup_{\mathbf{x}\in D}\Phi_{\mathbf{x}}\,.$$

Finally, observe that

$$\mathbb{E}\sup_{\mathbf{x}\in D}\Phi_{\mathbf{x}} = \mathbb{E}\sup_{\mathbf{x}\in D}\Psi_{\mathbf{x}}\,.$$

The lemma is obtained taking the infimum over $\mathbf{x}_0 \in D$. □

### 3.3.4 A Lower Bound in the Worst Case Setting

As mentioned in Section 3.2.3, commonly used discretization techniques are not feasible for tractability analysis. Osipenko and Parfenov 1995 [61], Kuo, Wasilkowski, and Woźniakowski 2008 [40], and Cobos, Kühn, and Sickel 2016 [12], independently from each other found similar approaches to relate the error of $L_\infty$-approximation to $L_2$-approximation. The formulation of Proposition 3.10 follows [12, 61], giving a lower bound for the $L_\infty$-approximation in terms of singular values of some $L_2(\rho)$-approximation. Parts of the proof in the original papers are based on the theory of absolutely summing operators. For this thesis, however, a proof that is based on tools from IBC appears more natural. Namely, in Kuo et al. [40] the worst case $L_\infty$ error was compared to the average $L_2(\rho)$ error with respect to the Gaussian field associated with the reproducing kernel Hilbert space $\mathcal{H} = \mathcal{H}(K)$.

We start with a well known fact on the optimality of linear algorithms for the approximation of Hilbert space functions. The proofs are given for completeness.

**Lemma 3.9.** *Consider a linear problem $S : \mathcal{H} \to G$ with the input set $F$ being the unit ball of a Hilbert space $\mathcal{H}$.*

*(a) Linear algorithms are optimal, more precisely, optimal algorithms have the structure $A_n = SP$ where $P$ is an orthogonal rank-$n$ projection on $\mathcal{H}$. Hence we can write*

$$e^{\mathrm{det}}(n, S, \Lambda^{\mathrm{all}}) = \inf_{\substack{P\ \mathit{Proj.}\\ \mathrm{rk}\,P=n}} \|S\,(\mathrm{id}_{\mathcal{H}} - P)\|_{\mathcal{H}\to G}\,.$$

*(b) If $G = \mathcal{H}_2$ is another Hilbert space and $S$ is compact, we have a singular value decomposition. That is, there is an orthonormal system $(\psi_k)_{k=1}^M$ in $\mathcal{H}$ such that $(S\psi_k)_{k=1}^M$ is orthogonal in $G = \mathcal{H}_2$, and $S$ can be written*

$$Sf = \sum_{k=1}^{M} \langle \psi_k, f \rangle_{\mathcal{H}} \, S\psi_k \,,$$

*where $M \in \mathbb{N} \cup \{\infty\}$. Furthermore, the sequence $(\sigma_k)_{k=1}^{\infty}$ of the singular values*

$$\sigma_k := \begin{cases} \|S\psi_k\|_G > 0 & \text{for } k < M+1, \\ 0 & \text{for } k > M, \end{cases}$$

*is ordered, $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$. Then*[15]

$$e^{\mathrm{det}}(n, S, \Lambda^{\mathrm{all}}) = \sigma_{n+1} \,.$$

*Proof.* (a) Let $N : \mathcal{H} \to \mathbb{R}^n$ be any deterministic information mapping using adaptively chosen functionals $L_{k,\mathbf{y}_{[k-1]}}$. We define the non-adaptive information

$$N^{\mathrm{non}}(f) := (L_1(f), L_{2,0}(f), L_{3,\mathbf{0}_{[2]}}(f), \ldots, L_{n,\mathbf{0}_{[n-1]}}(f)) \,,$$

which uses the functionals that $N$ would choose in the case of $N(f) = \mathbf{0}$. Let $P$ be the orthogonal projection onto $\ker(N^{\mathrm{non}})^{\perp}$, i.e. $\mathrm{id}_{\mathcal{H}} - P$ is the orthogonal projection onto $\ker(N^{\mathrm{non}})$, and consider the linear rank-$n$ algorithm $A_n := SP$ (it is indeed based on the information $N^{\mathrm{non}}$ since $\ker N^{\mathrm{non}} = \ker SP$). The error of this algorithm is

$$\begin{aligned} e(A_n, S) \;=\; \sup_{f \in F} \|S\,(\mathrm{id}_{\mathcal{H}} - P)\, f\|_G \;&=\; \sup_{\substack{f \in \mathrm{img}(\mathrm{id}_{\mathcal{H}} - P) \\ \|f\|_{\mathcal{H}} \leq 1}} \|S\,f\|_G \;=\; \sup_{\substack{f \in F \\ N(f) = \mathbf{0}}} \|S\,f\|_G \\ &\leq \inf_{\phi} \sup_{\substack{f \in F \\ N(f) = \mathbf{0}}} \|S\,f - [\phi \circ N](f)\|_G \;\leq\; \inf_{\phi} e(\phi \circ N, S) \,. \end{aligned}$$

This shows that the error of $A_n$ is maximal for zero information, a case which also occurs for the adaptive information mapping.

(b) The singular value decomposition is a standard result from spectral theory. For $n \geq M$, the statement is trivial since $S$ itself, with rank $\mathrm{rk}\, S \leq n$, can be seen as a suitable algorithm. Now, for $n < M$, take the rank-$n$ algorithm

$$A_n(f) = \sum_{k=1}^{n} \langle \psi_k, f \rangle_{\mathcal{H}} \psi_k \,,$$

[15] Following the axiomatic scheme of Pietsch [63], singular values of linear operators between Hilbert spaces are a special case of $s$-numbers, $s_n(S) := \sigma_n$. All kinds of $s$-numbers – which may differ for operators between arbitrary Banach spaces – coincide with the singular values in the Hilbert space setting. The correspondence to the error of deterministic methods exhibits the usual index shift we encounter when relating quantities from IBC to $s$-numbers, see also the discussion towards the end of Section 1.4, and the definition of Bernstein numbers in Section 2.1.

with the error

$$e(A_n, f) = \Big\| \sum_{k=n+1}^{M} \langle \psi_k, f\rangle_{\mathcal{H}}\, \psi_k \Big\|_G = \sqrt{\sum_{k=n+1}^{M} \langle \psi_k, f\rangle_{\mathcal{H}}^2\, \sigma_k^2}$$
$$\leq \sigma_{n+1} \sqrt{\sum_{k=n+1}^{M} \langle \psi_k, f\rangle_{\mathcal{H}}^2} \leq \sigma_{n+1}\, \|f\|_{\mathcal{H}}\,,$$

where equality is attained for $f = \psi_{n+1}$. This gives us the upper bound.
This upper bound is optimal. Indeed, for any rank-$n$ algorithm $A_n = S\,P$, there exists an element $f \in \operatorname{span}\{\psi_1, \ldots, \psi_{n+1}\}$ with $\|f\|_{\mathcal{H}} = 1$ and $A_n(f) = 0$, wherefore

$$e(A_n, f) = \|f\|_G = \sqrt{\sum_{k=1}^{n+1} \sigma_k^2\, \langle \psi_k, f\rangle_{\mathcal{H}}^2} \geq \sigma_{n+1} \|f\|_{\mathcal{H}}\,.$$

This implies the matching lower bound

$$e^{\mathrm{det}}(n, S, \Lambda^{\mathrm{all}}) \geq \sigma_{n+1}\,.$$

□

Let $\rho$ be a measure on $D$ (defined for Borel sets in $D$, with respect to the canonical metric $d_K$). Recall that $L_p(\rho)$ denotes the space of (equivalence classes of) measurable functions defined on $D$ and bounded in the norm

$$\|f\|_{L_p(\rho)} := \begin{cases} \left(\int_D f^p \,\mathrm{d}\rho\right)^{1/p} & \text{for } 1 \leq p < \infty, \\ \operatorname{ess\,sup}_{D,\rho} |f| & \text{for } p = \infty, \end{cases}$$

where $\operatorname{ess\,sup}_{D,\rho} |f| := \sup\{\lambda \in \mathbb{R} \mid \rho\{\mathbf{x} \in D : |f(\mathbf{x})| \geq \lambda\} > 0\}$. For continuous functions it makes sense to consider the supremum norm

$$\|f\|_{\sup} := \sup_{\mathbf{x} \in D} |f(\mathbf{x})| \geq \|f\|_{L_\infty(\rho)}\,,$$

later, when the supremum norm and the $L_\infty$-norm coincide, we will only write $\|\cdot\|_\infty$.

The following version of a worst case lower bound is close to the formulation of Osipenko and Parfenov [61, Thm 3], also Cobos et al. [12, Lem 3.3], however, it is essentially contained in Kuo et al. [40] as well.[16] The first part of the proof follows Kuo et al. [40, Thm 1].

**Proposition 3.10.** *Let $\rho$ be a probability measure on the domain $D$, and let the separable reproducing kernel Hilbert space $\mathcal{H} = \mathcal{H}(K)$ be compactly embedded into $L_\infty(\rho)$. The embedding $\mathcal{H} \hookrightarrow L_2(\rho)$ is compact as well, and a singular value decomposition exists. This means, there is an orthonormal basis $(\psi_k)_{k=1}^M$ (with $M \in \mathbb{N} \cup \{\infty\}$)*

[16] Kuo et al. in their research work with eigenvalues of an integral operator defined via the kernel function $K$. These eigenvalues are the squared singular values, which in turn we prefer to use here.

*of $\mathcal{H}$ which is orthogonal in $L_2(\rho)$ as well, and the corresponding singular values $\sigma_k := \|\psi_k\|_{L_2(\rho)}$, for $k < M+1$, are in decaying order $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$.*

*Then we have*

$$e^{\mathrm{det}}(n, \mathcal{H} \hookrightarrow L_\infty(\rho), \Lambda^{\mathrm{all}}) \geq \sqrt{\sum_{k=n+1}^{\infty} \sigma_k^2}\,.$$

*Proof.* Without loss of generality $n < M$.

By Lemma 3.9 we know that optimal algorithms with cardinality $n$ for the approximation problem $\mathrm{APP} : \mathcal{H} \hookrightarrow L_\infty(\rho)$ can be built with orthonormal $\varphi_1, \ldots, \varphi_n \in \mathcal{H}$,

$$A_n(f) = \sum_{j=1}^{n} \langle \varphi_j, f \rangle_{\mathcal{H}}\, \varphi_j\,.$$

The orthonormal system can be completed to an orthonormal basis $(\varphi_k)_{k=1}^{M}$ of $\mathcal{H}$. Then the worst case error is

$$e(A_n, \mathcal{H} \hookrightarrow L_\infty(\rho)) = \operatorname*{ess\,sup}_{\mathbf{x} \in D} \sqrt{\sum_{j=n+1}^{M} \varphi_j^2(\mathbf{x})} \geq \sqrt{\int \sum_{j=n+1}^{M} \varphi_j^2(\mathbf{x})\, \rho(\mathrm{d}\mathbf{x})}\,, \qquad (3.3.7)$$

compare (3.3.4).[17]

Consider the Gaussian field $\Psi$ associated to $\mathcal{H}$,

$$\Psi_{\mathbf{x}} = \sum_{j=1}^{M} X_j\, \varphi_j(\mathbf{x})\,,$$

where the $X_j$ are independent standard Gaussian random variables, and let $\mu$ denote the distribution of $\Psi$. The intention is to study the algorithm $A_n$ for the $L_2(\rho)$-approximation in the $\mu$-average setting. For $M = \infty$ however, $\Psi \notin \mathcal{H}$ almost surely, but $A_n$ uses functionals that are defined for functions from $\mathcal{H}$. So instead, consider the random functions $\Psi^{(m)}$ for $m \in \mathbb{N}$,

$$\Psi^{(m)}_{\mathbf{x}} := \sum_{j=1}^{m \wedge M} X_j\, \varphi_j(\mathbf{x})\,,$$

and let $\mu^{(m)}$ denote the corresponding distribution in $\mathcal{H}$. Clearly,

$$\Psi^{(m)}_{\mathbf{x}} - [A_n\, \Psi^{(m)}](\mathbf{x}) = \sum_{j=n+1}^{m \wedge M} X_j\, \varphi_j(\mathbf{x})\,,$$

[17] The proof is simpler once knowing that the optimal algorithm, in fact, is built of an orthogonal projection within $\mathcal{H}$ in composition with the solution operator. In Kuo et al. [40, Thm 1] a little more work is needed because it was only used that optimal algorithms are linear.

and for the root mean square average $L_2(\rho)$-error we have

$$e_2(A_n, L_2(\rho), \mu^{(m)}) = \sqrt{\mathbb{E} \int \left( \sum_{j=n+1}^{m \wedge M} X_j \, \varphi_j(\mathbf{x}) \right)^2 \rho(\mathrm{d}\mathbf{x})}$$

$$[\text{Fubini, } X_j \text{ iid}] \quad = \sqrt{\int \sum_{j=n+1}^{m \wedge M} \varphi_j^2(\mathbf{x}) \, \rho(\mathrm{d}\mathbf{x})} \, . \tag{3.3.8}$$

In comparison with (3.3.7), this shows

$$e(A_n, \mathcal{H} \hookrightarrow L_\infty(\rho)) \geq \lim_{m \to \infty} e_2(A_n, L_2(\rho), \mu^{(m)}) \, , \tag{3.3.9}$$

where the limit is approached monotonously from below.

The RHS of (3.3.9) can be expressed by means of singular values of the compact mapping $[\mathrm{id} - A_n] : \mathcal{H} \to L_2(\rho)$. In detail, there exists an orthonormal system $(\chi_i)_{i=1}^{M'}$ in $\mathcal{H}$ such that $([\mathrm{id} - A_n] \, \chi_i)_{i=1}^{M'}$ is orthogonal in $L_2(\rho)$ and

$$[\mathrm{id} - A_n] f = \sum_{i=1}^{M'} \langle \chi_i, f \rangle_{\mathcal{H}} \, [\mathrm{id} - A_n] \chi_i \, ,$$

with the singular values $\tau_i := \| [\mathrm{id} - A_n] \, \chi_i \|_{L_2(\rho)}$, for $i < M' + 1 := M - n + 1$, as always in decaying order $\tau_1 \geq \tau_2 \geq \ldots > 0$. (For $i > M'$ we have $\tau_i = 0$.) With

$$Z_i^{(m)} := \langle \chi_i, \Psi^{(m)} \rangle_{\mathcal{H}} = \sum_{j=1}^{m \wedge M} \langle \chi_i, \varphi_j \rangle_{\mathcal{H}} \, X_j \, ,$$

we can write

$$e_2(A_n, L_2(\rho), \mu^{(m)}) = \sqrt{\mathbb{E} \, \| [\mathrm{id} - A_n] \Psi^{(m)} \|_{L_2(\rho)}^2} = \sqrt{\mathbb{E} \left\| \sum_{i=1}^{M'} Z_i^{(m)} \, [\mathrm{id} - A_n] \chi_i \right\|_{L_2(\rho)}^2}$$

$$= \sqrt{\sum_{i=1}^{M'} \tau_i^2 \, \mathbb{E}(Z_i^{(m)})^2} \, .$$

Note that

$$\mathbb{E}(Z_i^{(m)})^2 \overset{[X_j \text{ iid}]}{=} \sum_{j=1}^{m \wedge M} \langle \chi_i, \varphi_j \rangle_{\mathcal{H}}^2 \xrightarrow[m \to \infty]{} \| \chi_i \|_{\mathcal{H}}^2 = 1 \, ,$$

where the sequence is monotonically increasing in $m$, so we have

$$\lim_{m \to \infty} e_2(A_n, L_2(\rho), \mu^{(m)}) = \sqrt{\sum_{i=1}^{\infty} \tau_i^2} \, .$$

It remains to show that $\tau_i \geq \sigma_{n+i}$, and by (3.3.9) we are done. Consider any algorithm $A'_m$ for the approximation of $[\mathrm{id} - A_n] : \mathcal{H} \to L_2(\rho)$. Then $[A_n + A'_{i-1}]$ is

an algorithm of cardinality $(n+i-1)$ for the approximation of $\mathrm{id} : \mathcal{H} \hookrightarrow L_2(\rho)$. By this observation and Lemma 3.9 (b), we obtain

$$\begin{aligned}
\tau_i &= e^{\det}(i-1, [\mathrm{id} - A_n] : \mathcal{H} \to L_2(\rho)) \\
&= \inf_{A'_{i-1}} \sup_{\|f\|_{\mathcal{H}} \le 1} \|[\mathrm{id} - A_n] f - A'_{i-1} f\|_{L_2(\rho)} \\
&\ge \inf_{A''_{n+i-1}} \sup_{\|f\|_{\mathcal{H}} \le 1} \|f - A''_{n+i-1} f\| \\
&= e^{\det}(n+i-1, \mathrm{id} : \mathcal{H} \hookrightarrow L_2(\rho)) \\
&= \sigma_{n+i} \,.
\end{aligned}$$

□

**Example 3.11** (Diagonal operators)**.** The proposition above can be used to prove lower bounds for the approximation of diagonal operators on sequence spaces. This has already been pointed out by Osipenko and Parfenov [61, Sec 4]. We repeat the example for its connection to the RKHS framework.

Consider a compact matrix operator

$$A : \ell_2 \to \ell_\infty, \quad (x_j)_{j \in \mathbb{N}} \mapsto \left( \textstyle\sum_j a_{ij}\, x_j \right)_{i \in \mathbb{N}} .$$

This problem is equivalent to the embedding operator

$$\mathrm{id} : \mathcal{H}(K) \hookrightarrow \ell_\infty \,,$$

where we have the reproducing kernel

$$K : \mathbb{N} \times \mathbb{N} \to \mathbb{R}, \quad K(i,j) = (A\, A^\top)(i,j) \,.$$

Let $\rho$ be a probability measure on $\mathbb{N}$, with $\rho_i \ge 0$ denoting the probability of $i \in \mathbb{N}$, we have $\sum_{i=1}^\infty \rho_i = 1$. Then $\ell_2(\rho)$ is the space of sequences $\mathbf{x} \in \mathbb{R}^{\mathbb{N}}$ that are bounded in the norm

$$\|\mathbf{x}\|_{\ell_2(\rho)} := \sqrt{\sum_{i=1}^\infty \rho_i\, x_i^2} \,.$$

Now, let $A$ be a diagonal operator

$$[A\mathbf{x}](i) = \lambda_i\, x_i \quad \text{for } i \in \mathbb{N},$$

where $\lambda_1 \ge \lambda_2 \ge \ldots \ge 0$ and $\lambda_i \xrightarrow[i \to \infty]{} 0$. The reproducing kernel for the Hilbert space of the corresponding embedding problem is

$$K(i,j) = \delta_{ij}\, \lambda_i^2 \,.$$

Hence $(\lambda_i \mathbf{e}_i)_{i=1}^M$ is an orthonormal basis of $\mathcal{H}(K)$, here $M := \inf\{m \in \mathbb{N}_0 \mid \lambda_{m+1} = 0\}$. It is also orthogonal in $\ell_2(\rho)$ for any measure $\rho$ on $\mathbb{N}$. Taking

$$\rho_i := \frac{\lambda_i^{-2}}{\sum_{j=1}^m \lambda_j^{-2}}\, \mathbb{1}[i \le m] \,,$$

with $1 \le m < M+1$, for $\mathcal{H}(K) \hookrightarrow \ell_2(\rho)$ we have the singular values

$$\sigma_1 = \ldots = \sigma_m = \left( \sum_{j=1}^{m} \lambda_j^{-2} \right)^{-\frac{1}{2}} > 0 = \sigma_{m+1} = \sigma_{m+2} = \ldots$$

By Proposition 3.10 we obtain

$$e^{\mathrm{det}}(n, A : \ell_2 \to \ell_\infty, \Lambda^{\mathrm{all}}) \ge \sqrt{\frac{m-n}{\sum_{j=1}^{m} \lambda_j^{-2}}},$$

taking the supremum over $m$, $n \le m < M+1$, this gives sharp lower bounds.

The proof for the lower bound does not reveal any information about the structure of optimal methods. A construction of methods can be found in the book of Osipenko [60, pp. 155-159]. The finite dimensional case goes back to Smolyak 1965 [72], the general case to Hutton, Morrell, and Retherford [29, Thm 2.12].

## 3.4 Breaking the Curse for Function Approximation

We study two examples of tensor product Hilbert spaces where for $L_\infty$-Approximation we can show the curse of dimensionality for the worst case, but in the randomized setting we have polynomial tractability with $n^{\mathrm{ran}}(\varepsilon, d) \preceq d\,(1+\log d)/\varepsilon^2$. The first example on the approximation with the Brownian sheet, see Section 3.4.1, is more or less a toy example intended to demonstrate the new techniques in a setting that is easy to visualize, read Remark 3.14 for further comments on the rating of this example. For the second example on unweighted periodic tensor product Hilbert spaces, see Section 3.4.2, we find some general conditions that are sufficient for Monte Carlo to break the curse. These conditions are specified for Korobov spaces, see Theorem 3.19. The latter constitutes the main application of the present chapter. We will close the chapter with some final hints that one should bear in mind when searching for further examples, see Section 3.4.3.

### 3.4.1 Approximation with the Brownian Sheet

We consider the Hilbert space $\mathcal{H}(K_d)$ with the *Wiener sheet kernel* $K_d$ on $[-1,+1]^d$, the associated continuous Gaussian field $W$ is also called *Brownian sheet*, see e.g. Adler [1, p. 7, pp. 68/69] for basic properties of this space and the associated Gaussian random field.[18]

For $d = 1$ we take the covariance kernel of the "two-armed" Brownian motion,

$$\begin{aligned} K_1(x,z) &:= \frac{|x| + |z| - |x-z|}{2} \\ &= \mathbb{1}[\operatorname{sgn} x = \operatorname{sgn} z]\, \min\{|x|, |z|\}\,. \end{aligned}$$

[18] In the lecture notes of Adler, sometimes a *set-indexed Brownian sheet* is dealt with. We do not need that concept here.

This space consists of once weakly differentiable functions,

$$\mathcal{H}(K_1) = \left\{ f : [-1,+1] \to \mathbb{R} \mid f(0) = 0, \|f\|_{\mathcal{H}(K_1)} := \|f'\|_2 \right\} ,$$

the inner product is $\langle f, g \rangle_{\mathcal{H}(K_1)} := \int_{-1}^{+1} f'(x)\, g'(x)\, \mathrm{d}x$.

For $d \in \mathbb{N}$ we take the tensor product kernel,

$$\begin{aligned} K_d(\mathbf{x}, \mathbf{z}) &:= \prod_{j=1}^{d} K_1(x_j, z_j) \\ &= \mathbb{1}[\operatorname{sgn} \mathbf{x} = \operatorname{sgn} \mathbf{z}] \prod_{j=1}^{d} \min\{|x_j|, |z_j|\} , \end{aligned}$$

where $\operatorname{sgn} \mathbf{x} = (\operatorname{sgn} x_j)_{j=1}^d$ is the vector-valued signum function. Thus

$$\begin{aligned} \mathcal{H}(K_d) &= \bigotimes_{j=1}^{d} \mathcal{H}(K_1) \\ &= \{ f : [-1,+1]^d \to \mathbb{R} \mid f(\mathbf{x}) = 0 \text{ if } \exists_j\, x_j = 0, \text{ and } \|f\|_{\mathcal{H}} := \|D^{\mathbf{1}} f\|_{L_2} < \infty \} , \end{aligned}$$

where $D^{\mathbf{1}}(f) := \partial_1 \cdots \partial_d f$. Note that the initial error is 1 (thus the problem is properly normalized) since the kernel takes its maximum 1 in the corners $\{\pm 1\}^d$. Furthermore, functions $f \in \mathcal{H}(K_d)$ can be identified with functions $\tilde{f} \in \mathcal{H}(K_{d+1})$,

$$\tilde{f}(x_1, \ldots, x_{d+1}) := f(x_1, \ldots, x_d)\, [x_{d+1}]_+ ,$$

then $\|f\|_{\mathcal{H}} = \|\tilde{f}\|_{\mathcal{H}}$, and $\|f\|_\infty = \|\tilde{f}\|_\infty$, where for $\tilde{f}$ the maximal absolute value is attained with $x_{d+1} = 1$.

**Theorem 3.12.** *Deterministic $L_\infty$-approximation of functions from the Wiener sheet space on $[-1,+1]^d$ suffers from the curse of dimensionality, in detail,*

$$n^{\mathrm{det}}(\varepsilon, \mathcal{H}(K_d) \hookrightarrow L_\infty, \Lambda^{\mathrm{all}}) \geq 2^{d-1} , \quad \textit{for } 0 < \varepsilon \leq \frac{\sqrt{2}}{2}.$$

*Proof.* We consider the $2^d$-dimensional subspace of $\mathcal{H}(K_d)$ spanned by $\{K_{\boldsymbol{\sigma}}\}_{\boldsymbol{\sigma} \in \{\pm 1\}^d}$, where

$$K_{\boldsymbol{\sigma}}(\mathbf{x}) \;:=\; K_d(\boldsymbol{\sigma}, \mathbf{x}) \;=\; \mathbb{1}[\operatorname{sgn} \mathbf{x} = \boldsymbol{\sigma}] \prod_{j=1}^{d} |x_j| .$$

These functions are orthonormal in $\mathcal{H} = \mathcal{H}(K_d)$ since

$$\langle K_{\boldsymbol{\sigma}}, K_{\boldsymbol{\sigma}'} \rangle_{\mathcal{H}} \;=\; K_d(\boldsymbol{\sigma}, \boldsymbol{\sigma}') \;=\; \mathbb{1}[\boldsymbol{\sigma} = \boldsymbol{\sigma}'] .$$

Besides, they have essentially disjoint supports (which are the subcubes of $[-1,+1]^d$ with constant sign in each coordinate), and take their supremum in $\mathbf{x} = \boldsymbol{\sigma}$, which is $K_{\boldsymbol{\sigma}}(\boldsymbol{\sigma}) = 1$. For $f = \sum_{\boldsymbol{\sigma} \in \{\pm 1\}^d} a_{\boldsymbol{\sigma}}\, K_{\boldsymbol{\sigma}}$ we have

$$\|f\|_{\mathcal{H}(K_d)} = \sqrt{\sum_{\boldsymbol{\sigma} \in \{\pm 1\}^d} a_{\boldsymbol{\sigma}}^2} , \quad \text{and} \quad \|f\|_{L_\infty} = \max_{\boldsymbol{\sigma} \in \{\pm 1\}^d} |a_{\boldsymbol{\sigma}}| .$$

Hence we can estimate the error from below by comparison to a sequence space embedding,

$$e^{\mathrm{det}}(n, \mathcal{H}(K_d) \hookrightarrow L_\infty) \geq e^{\mathrm{det}}(n, \ell_2^{2^d} \hookrightarrow \ell_\infty^{2^d}) \geq \sqrt{1 - n\, 2^{-d}}\,,$$

see (3.2.4). This implies the stated lower bound for the complexity, compare Section 3.2.4. □

**Theorem 3.13.** *Randomized $L_\infty$-approximation of functions from the Wiener sheet space, using linear functionals for information, is polynomially tractable. In detail,*

$$n^{\mathrm{ran}}(\varepsilon, \mathcal{H}(K_d) \hookrightarrow L_\infty, \Lambda^{\mathrm{all}}) \leq C\, \frac{d\,(1 + \log d)}{\varepsilon^2}\,,$$

*with a numerical constant $C > 0$.*

*Proof.* We want to apply Lemma 3.3, so we need to estimate

$$\mathbb{E}\, \|W\|_\infty$$

for the Brownian sheet $W$ on $[-1, +1]^d$ defined by the covariance kernel $K_d$. This will be done by an entropy estimate and Proposition 3.7 (Dudley).[19]

The canonical metric of the Wiener sheet kernel is

$$d_K(\mathbf{x}, \mathbf{z})^2 = \prod_{j=1}^d |x_j| + \prod_{j=1}^d |z_j| - 2\, \mathbb{1}[\operatorname{sgn} \mathbf{x} = \operatorname{sgn} \mathbf{z}] \left( \prod_{j=1}^d \min\{|x_j|, |z_j|\} \right).$$

If $\operatorname{sgn} \mathbf{x} \neq \operatorname{sgn} \mathbf{z}$ there exists an index $j \in \{1, \ldots, d\}$ such that $\operatorname{sgn} x_j \neq \operatorname{sgn} z_j$, hence

$$d_K(\mathbf{x}, \mathbf{z})^2 = \prod_{j=1}^d |x_j| + \prod_{j=1}^d |z_j| \leq |x_j| + |z_j| = |x_j - z_j| \leq |\mathbf{x} - \mathbf{z}|_\infty\,.$$

[19] The same approach for estimating the expected maximum of $W$ is taken in Adler [1, Prop 1.2] though for the domain $[0, 1]^d$ instead, which admittedly is a minor change. Steps that have been left to the reader there are explicated here.

If $\operatorname{sgn}\mathbf{x} = \operatorname{sgn}\mathbf{z}$, we obtain

$$\begin{aligned}
d_K(\mathbf{x},\mathbf{z})^2 &= \prod_{j=1}^d |x_j| + \prod_{j=1}^d |z_j| - 2\prod_{j=1}^d \min\{|x_j|,|z_j|\} \\
&\le 2\left(\prod_{j=1}^d \max\{|x_j|,|z_j|\} - \prod_{j=1}^d \min\{|x_j|,|z_j|\}\right) \\
\text{[telescoping sum]} \quad &\le 2\sum_{k=1}^d \Bigg[\underbrace{\left(\prod_{j=1}^{k-1} \min\{|x_j|,|z_j|\}\right)}_{\le 1} \\
&\qquad \underbrace{(\max\{|x_k|,|z_k|\} - \min\{|x_k|,|z_k|\})}_{=|x_k-z_k|} \\
&\qquad \underbrace{\left(\prod_{j=k+1}^d \max\{|x_j|,|z_j|\}\right)}_{\le 1}\Bigg] \\
&\le 2\,|\mathbf{x}-\mathbf{z}|_1\,.
\end{aligned}$$

This shows

$$d_K(\mathbf{x},\mathbf{z})^2 \le 2\,d\,|\mathbf{x}-\mathbf{z}|_\infty\,,$$

and for $\mathbf{x} \in [-1,1]^d$ and $r > 0$ we have the inclusion

$$B_\infty\left(\mathbf{x}, \frac{r^2}{2\,d}\right) \subseteq B_K(\mathbf{x}, r)\,,$$

where $B_\infty$ denotes the ball in the $\ell_\infty^d$-metric. Since one can cover $B_\infty(\mathbf{0},1) = [-1,1]^d$ by $\lceil 2\,d/r^2\rceil^d$ balls with $\ell_\infty^d$-radius $\frac{r^2}{2\,d}$, for the metric entropy with respect to $d_K$ we obtain

$$H(r) = \log(N(r)) \le d\log\left(1 + \frac{2\,d}{r^2}\right) \le C_1\,d\,(1+\log d)\,(1-\log r)\,.$$

Note that $N(1) = 1$ with the ball around $\mathbf{x} = \mathbf{0}$.

Since the Brownian sheet $W$ is zero on the coordinate hyperplanes, we do not need the additional term when applying Lemma 3.8,

$$\begin{aligned}
\mathbb{E}\,\|W\|_\infty &\le 2\,\mathbb{E}\sup_{\mathbf{x}\in[-1,1]^d} W_{\mathbf{x}} \\
\text{[Proposition 3.7]} \quad &\le 2\,C_{\text{Dudley}}\,C_1\sqrt{d\,(1+\log d)}\int_0^1 \sqrt{1-\log r}\,\mathrm{d}r \\
\text{[subst. } \tfrac{s^2}{2} = 1-\log r\text{]} \quad &= C_2\sqrt{d\,(1+\log d)}\,\frac{\mathrm{e}}{\sqrt{2}}\int_1^\infty s^2\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s] \\
&= C_3\sqrt{d\,(1+\log d)}\,.
\end{aligned}$$

Hence by Lemma 3.3,

$$n^{\mathrm{ran}}(\varepsilon, \mathcal{H}(K_d) \hookrightarrow L_\infty, \Lambda^{\mathrm{all}}) \leq 4\, C_3^2\, \frac{d\,(1+\log d)}{\varepsilon^2}\,.$$

This finishes the proof. □

**Remark 3.14.** This is a first example of a function approximation problem where Monte Carlo methods can break the curse of dimensionality. As mentioned before, the initial error is properly normalized, and the problem for lower dimensions is contained in the problem for higher dimensions. This example, however, has a downside: Actually, we treat the simultaneous approximation of $2^d$ entirely independent functions (that only need to be bounded in a common Euclidean norm). This view is justified by the fact that functions from the space $\mathcal{H}(K_d)$ are zero at the coordinate hyperplanes, that way the domain is split into subcubes of constant sign. Observe the similarities with the sequence space example in Section 3.2.4 where we only lack the logarithmic term $(1+\log d)$ in the Monte Carlo upper bound. Adding some proper "function space nature" by this Wiener sheet example, honestly, serves as a fig-leaf for the artificiality of the sequence space example. The next section treats much more natural problems.

Including this example in this study, however, was motivated by the fact that the Brownian sheet is widely known, and that the loss of smoothness becomes palpable. In addition, this gives us a non-periodic example where it was convenient to use entropy methods for the estimate (in contrast to the next section where we will rely on the technique of majorizing measures). The Wiener sheet – usually only defined on $[0,1]^d$ – is a common example for many topics in IBC, see for example Novak and Woźniakowski [55, 57, 58] or Ritter [68].

### 3.4.2 Tensor Product Spaces of Periodic Functions

#### The General Setting

We consider the $L_\infty$-approximation of Hilbert space functions defined on the $d$-dimensional torus $\mathbb{T}^d$, compare the notation in Cobos et al. [12] (with slight modifications). The Hilbert spaces we consider will be unweighted tensor product spaces.

A few words on the domain. The one-dimensional torus $\mathbb{T} := \mathbb{R} \bmod \mathbb{Z} \equiv [0,1)$ can be identified with the unit interval tying the endpoints together. A natural way to define a metric on $\mathbb{T}$ is

$$d_{\mathbb{T}}(x,z) := \min_{k\in\{-1,0,1\}} |x - z + k|\,, \quad \text{for } x, z \in [0,1).$$

This is the length of the shortest connection between two points along a closed curve of length 1. For the $d$-dimensional torus we take the summing metric

$$d_{\mathbb{T}^d}(\mathbf{x},\mathbf{z}) := \sum_{j=1}^{d} d_{\mathbb{T}}(x_j, z_j)\,.$$

Smoothness and continuity are to be defined with respect to this metric.

We start with a basis representation of spaces under consideration. First, for $d = 1$, the Fourier system

$$\{\varphi_0 := 1,\, \varphi_{-k} := \sqrt{2}\,\sin(2\pi k\,\cdot),\, \varphi_k := \sqrt{2}\,\cos(2\pi k\,\cdot)\}_{k\in\mathbb{N}}$$

is an orthonormal basis for $L_2(\mathbb{T}) = L_2([0,1))$. We consider Hilbert spaces where these functions are still orthogonal. Namely, let $\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T})$ denote the Hilbert space for which the system

$$\{\psi_0 := \lambda_0,\, \psi_{-k} := \lambda_k\,\sin(2\pi k\,\cdot),\, \psi_k := \lambda_k\,\cos(2\pi k\,\cdot)\}_{k\in\mathbb{N}} \setminus \{0\}\,,$$

is an orthonormal basis.[20] Here, $\boldsymbol{\lambda} = (\lambda_k)_{k\in\mathbb{N}_0} \subset [0,\infty)$ indicates the importance of the different frequencies. Now, for general $d \in \mathbb{N}$, we consider the *unweighted*[21] tensor product space $\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d)$ with the tensor product orthonormal basis $\{\psi_{\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d} \setminus \{0\}$,

$$\psi_{\mathbf{k}}(\mathbf{x}) := \prod_{j=1}^{d} \psi_{k_j}(x_j)\,. \tag{3.4.1}$$

Analogously, we write $\{\varphi_{\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$ for the Fourier basis of $L_2(\mathbb{T}^d)$, once more at the risk of some confusion from using the same letter for the index as in the one-dimensional case, merely with a different font style.

For a suitable choice of the $\lambda_k$, we have the one-dimensional reproducing kernel

$$\begin{aligned} K_{\boldsymbol{\lambda}}(x,z) &:= \lambda_0^2 + \sum_{k=1}^{\infty} \lambda_k^2\,[\cos(2\pi k\,x)\,\cos(2\pi k\,z) + \sin(2\pi k\,x)\,\sin(2\pi k\,z)] \\ &= \sum_{k=0}^{\infty} \lambda_k^2\,\cos(2\pi k\,(x-z))\,, \end{aligned} \tag{3.4.2}$$

for general dimensions $d \in \mathbb{N}$ we obtain the product kernel

$$K_{\boldsymbol{\lambda}}^d(\mathbf{x},\mathbf{z}) := \prod_{j=1}^{d} K_{\boldsymbol{\lambda}}(x_j, z_j)\,.$$

In particular, the initial error is

$$e(0, \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)) = \sup_{\mathbf{x}\in\mathbb{T}^d} \sqrt{K_{\boldsymbol{\lambda}}^d(\mathbf{x},\mathbf{x})} = \left(\sum_{k=0}^{\infty} \lambda_k^2\right)^{d/2}\,.$$

The condition $\sum_{k=0}^{\infty} \lambda_k^2 < \infty$ is necessary and sufficient for the existence of a reproducing kernel and for the embedding $\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d) \hookrightarrow L_\infty$ to be compact, see Cobos et

[20] If $\lambda_k = 0$ for certain $k$, of course, the corresponding zero-functions cannot be part of the orthonormal basis. In the proof of Theorem 3.19 we consider finite-dimensional subspaces where the corresponding orthonormal basis will be $\{\psi_k\}_{k=-m}^{m}$. The same holds for the basis of the $d$-dimensional space.

[21] This means that every coordinate is equally important. For weighted tensor product spaces one would take different values for $\boldsymbol{\lambda}$ for different dimensions $j = 1, \ldots, d$, compare Cobos et al. [12], or Kuo et al. [40].

al. [12, Thm 3.1] with an extended list of equivalent properties. We will assume $\sum_{k=0}^{\infty} \lambda_k^2 = 1$ so that the initial error is constant 1.

Note that under this last assumption, functions $f \in \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d)$ can be identified with certain functions $\tilde{f} \in \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^{d+1})$ for $d < \tilde{d}$,

$$\tilde{f}(x_1, \dots, x_{\tilde{d}}) := f(x_1, \dots, x_d)\, K(0, x_{d+1})\,,$$

the $\mathcal{H}_{\boldsymbol{\lambda}}$- and the $L_\infty$-norms coincide, the maximum values of the function being attained for $x_{d+1} = 0$. So indeed, the problems of lower dimensions are contained in the problems of higher dimensions, yet $\tilde{f}$ is a bit lopsided in the redundant variable.

**Theorem 3.15.** *Suppose that $0 \le \lambda_0 < 1$ and $\sum_{k=0}^{\infty} \lambda_k^2 = 1$ for non-negative $\lambda_k$. Then the approximation problem*

$$\mathrm{APP} : \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)$$

*suffers from the curse of dimensionality in the deterministic setting.*

*In detail, while the initial error is constant 1, we have*

$$e^{\mathrm{det}}(n, \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)) \ge \sqrt{(1 - n\,\beta^d)_+}\,,$$

*where $\beta := \sup\{\lambda_0^2,\, \lambda_k^2/2\}_{k\in\mathbb{N}} \in (0,1)$. In other words, for $\varepsilon \in (0,1)$ we have the complexity bound*

$$n^{\mathrm{det}}(\varepsilon, \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)) \ge \beta^{-d}(1 - \varepsilon^2)\,.$$

*Proof.* Following Proposition 3.10, we study the singular values of $\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d) \hookrightarrow L_2(\mathbb{T}^d)$. Essentially, this can be traced back to the one dimensional case,

$$\psi_k = \sigma_k \varphi_k \quad \text{for } k \in \mathbb{Z},$$

where $\sigma_0 = \lambda_0$ and $\sigma_k = \sigma_{-k} = \lambda_k/\sqrt{2}$ for $k \in \mathbb{N}$ denote the unordered singular values of $\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}) \hookrightarrow L_2(\mathbb{T})$. In the multi-dimensional case we have

$$\psi_{\mathbf{k}} = \sigma_{\mathbf{k}} \varphi_{\mathbf{k}} \quad \text{for } \mathbf{k} \in \mathbb{Z}^d,$$

with the unordered singular values $\sigma_{\mathbf{k}} = \prod_{j=1}^d \sigma_{k_j}$, in particular

$$\sigma_{\mathbf{k}}^2 \le \left(\sup_{k'\in\mathbb{Z}} \sigma_{k'}^2\right)^d = \left(\sup\{\lambda_0^2,\, \lambda_{k'}^2/2\}_{k'\in\mathbb{N}}\right)^d = \beta^d\,.$$

On the other hand,

$$\sum_{\mathbf{k}\in\mathbb{Z}^d} \sigma_{\mathbf{k}}^2 = \left(\sum_{k'\in\mathbb{Z}} \sigma_{k'}^2\right)^d = \left(\sum_{k'=0}^{\infty} \lambda_{k'}^2\right)^d = 1\,.$$

So for any index set $I \subset \mathbb{Z}^d$ of size $\#I = n$, we have

$$\sqrt{\sum_{\mathbf{k}\in\mathbb{Z}^d\setminus I} \sigma_{\mathbf{k}}^2} \ge \sqrt{(1 - n\beta^d)_+}\,.$$

By Proposition 3.10, this proves the lower bound. □

**Remark 3.16.** Within the above proof we applied Proposition 3.10 with $\rho$ being the uniform distribution on $\mathbb{T}^d$. If we consider complex-valued Hilbert spaces, this approach will always give sharp lower bounds, see Cobos et al. [12, Thm 3.4]. In the real-valued setting we obtain sharp error results at least for those $n$ where the optimal index set $I \subset \mathbb{Z}^d$ contains all indices belonging to the same frequency, that is,

$$\mathbf{k} = (k_1, \dots, k_d) \in I \quad \Leftrightarrow \quad \operatorname{abs} \mathbf{k} := (|k_1|, \dots, |k_d|) \in I \,.$$

Still, in most cases it is hard to estimate the number of singular values within a certain range.

The following abstract result relies on estimates for the shape of the kernel function $K_{\mathbf{x}} = K(\mathbf{x}, \cdot)$.

**Theorem 3.17.** *Consider the uniform approximation problem*

$$\mathrm{APP} : \mathcal{H}(K_d) \hookrightarrow L_\infty(\mathbb{T}^d)$$

*where $\mathcal{H}(K_d)$ is a reproducing kernel Hilbert space on the $d$-dimensional torus $\mathbb{T}^d$ with the following properties:*

*(i) $K_d$ is the unweighted product kernel built from the one-dimensional case, this means $K_d(\mathbf{x}, \mathbf{z}) := \prod_{j=1}^d K_1(x_j, z_j)$ for $\mathbf{x}, \mathbf{z} \in \mathbb{T}^d$.*

*(ii) $K_1(x, x) = 1$ for all $x \in \mathbb{T}$.*
*(Consequently, $K_d(\mathbf{x}, \mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbb{T}^d$, in particular the initial error is constant 1.)*

*(iii) The kernel function can be locally estimated from below with an exponential decay, that is, there exist $\alpha > 0$ and $0 < R_0 \leq \frac{1}{2}$ such that*

$$K_1(x, z) \geq \exp(-\alpha \, d_{\mathbb{T}}(x, z)) \quad \text{for } x, z \in \mathbb{T} \text{ with } d_{\mathbb{T}}(x, z) \leq R_0.$$

*(Hence $K_d(\mathbf{x}, \mathbf{z}) \geq \exp(-\alpha \, d_{\mathbb{T}^d}(\mathbf{x}, \mathbf{z}))$ for $\max_j d_{\mathbb{T}}(x_j, z_j) < R_0$.)*[22]

*Then the problem is polynomially tractable in the randomized setting with general linear information $\Lambda^{\mathrm{all}}$, in detail,*

$$n^{\mathrm{ran}}(\varepsilon, \mathcal{H}(K_d) \hookrightarrow L_\infty(\mathbb{T}^d), \Lambda^{\mathrm{all}}) \leq C \, (1 + \alpha^2 - \log 2R_0) \, \frac{d \, (1 + \log d)}{\varepsilon^2} \,,$$

*with a universal constant $C > 0$.*

[22] If one is interested in a version of this theorem with better constants for particularly nice kernels, one could start with a stronger assumption $K_1(x, z) \geq \exp(-\alpha \, d_{\mathbb{T}}(x, z)^2)$ which is a comparison to a bell-shaped curve. The asymptotics of the complexity result, however, will not change. See also Remark 3.21 with a proposal for a modified version of this theorem.

*Proof.* We are going to apply the method of majorizing measures in order to estimate the expected maximum norm of the Gaussian field $\Psi$ associated with the reproducing kernel $K_d$. The majorizing measure $\mu$ we choose shall be the uniform distribution on $\mathbb{T}^d$, this is the Lebesgue measure on $[0,1)^d$.

Supposed that $\max_j d_{\mathbb{T}}(x_j, z_j) < R_0$, for the canonical metric in the $d$-dimensional case we have

$$\begin{aligned} d_K(\mathbf{x},\mathbf{z})^2 &= K_d(\mathbf{x},\mathbf{x}) + K_d(\mathbf{z},\mathbf{z}) - 2\,K_d(\mathbf{x},\mathbf{z}) \\ &\le 2\,(1-\exp(-\alpha\, d_{\mathbb{T}^d}(\mathbf{x},\mathbf{z}))) \\ &\le 2\,\alpha\, d_{\mathbb{T}^d}(\mathbf{x},\mathbf{z})\,. \end{aligned}$$

By this, we have the inclusion

$$B_{\mathbb{T}}\left(\sqrt{\frac{r}{2\,\alpha}},\,\mathbf{x}\right) \subseteq B_K(r,\mathbf{x})\,,$$

where $B_{\mathbb{T}}(R,\mathbf{x})$ denotes the $d_{\mathbb{T}^d}$-ball of radius $R$ around $\mathbf{x}\in\mathbb{T}^d$, and $B_K(r,\mathbf{x})$ is the ball of radius $r$ in the canonical metric associated with $K_d$. Hence

$$\mu(B_K(r,\mathbf{x})) \ge \mu\left(B_{\mathbb{T}}\left(\sqrt{\frac{r}{2\,\alpha}},\,\mathbf{x}\right)\right)\,.$$

We distinguish three cases:

- For $0 \le R \le R_0 \le \frac{1}{2}$, the $\mu$-volume $\mu(B_{\mathbb{T}}(R,\mathbf{x}))$ of the torus metric ball is the volume $\mathrm{Vol}(R\,B_1^d)$ of an $\ell_1$-ball in $\mathbb{R}^d$ with radius $R$, so with Stirling's formula,

$$\begin{aligned} \log(1/\mu(B_{\mathbb{T}}(R,\mathbf{x}))) &= \log(1/\,\mathrm{Vol}(R\,B_1^d)) \\ &= \log\Gamma(d+1) - d\,\log 2R \\ &\le C_1\,d\,(1+\log d)\,(1-\log 2R) \\ &\le C_1\,d\,(1+\log d)\,(1-\log(2R)^2)\,. \end{aligned} \tag{3.4.3}$$

  Such an estimate can be used for $0 \le r \le 2\alpha\,R_0^2 \le \alpha/2$.

- For $R > R_0$, the $\mu$-volume of $B_{\mathbb{T}}(R,\mathbf{x})$ can be estimated from below with the $\mu$-volume of an $\ell_1$-ball with radius $R_0 \le \frac{1}{2}$. We will use this in the case $2\alpha\,R_0^2 < r < 2$.

- For $r \ge 2$, we know $B_K(r,\mathbf{x}) = \mathbb{T}^d$ with $\mu$-volume 1 since $d_K(\mathbf{x},\mathbf{z}) \le 2$. In this case the term $\log(1/\mu(B_K(r,\mathbf{x})))$ vanishes.

Combining these cases, we can estimate

$$\begin{aligned}\int_0^\infty \sqrt{\log(1/\mu(B_K(\mathbf{x},r)))}\,\mathrm{d}r &\le \int_0^{\alpha/2} \sqrt{\log\left(1\Big/\mathrm{Vol}\left(\sqrt{\frac{r}{2\,\alpha}}\,B_1^d\right)\right)}\,\mathrm{d}r \\ &\quad + \left(2-2\alpha\,R_0^2\right)_+ \sqrt{\log\left(1/\mathrm{Vol}\left(R_0\,B_1^d\right)\right)} \\ &\overset{(3.4.3)}{\le} \sqrt{C_1\,d\,(1+\log d)} \\ &\qquad \left(\int_0^{\alpha/2} \sqrt{1-\log\frac{2r}{\alpha}}\,\mathrm{d}r \;+\; 2\,\sqrt{1-\log 2R_0}\right) \\ &= C_2\,(1+\alpha+\sqrt{-\log 2R_0})\,\sqrt{d\,(1+\log d)}\,.\end{aligned}$$

Here, the last integral can be transformed into a familiar integral by the substitution $s^2/2 = 1-\log\frac{2r}{\alpha}$,

$$\int_0^{\alpha/2} \sqrt{1-\log\frac{2r}{\alpha}}\,\mathrm{d}r = \frac{\mathrm{e}\,\alpha}{2\,\sqrt{2}}\int_1^\infty s^2\,\exp\left(-\frac{s^2}{2}\right)\,\mathrm{d}s\,.$$

Now, consider the Gaussian field $\Psi$ associated with the reproducing kernel $K_d$. Putting the above calculation into Proposition 3.6 (Fernique), with Lemma 3.8 we obtain

$$\begin{aligned}\mathbb{E}\,\|\Psi\|_\infty &\le \sqrt{\frac{2}{\pi}} + 2\,C_{\text{Fernique}}\,C_2\,(1+\alpha+\sqrt{-\log 2R_0})\,\sqrt{d\,(1+\log d)} \\ &\le C_3\,(1+\alpha+\sqrt{-\log 2R_0})\,\sqrt{d\,(1+\log d)}\,.\end{aligned}$$

By Lemma 3.3, this gives us a final upper bound on the complexity. □

We finish the general part of the periodic setting with sufficient conditions for the parameters $\boldsymbol{\lambda}$ of the kernels $K_{\boldsymbol{\lambda}}$ for that we can apply Theorem 3.17.

**Corollary 3.18.** *Given a kernel*

$$K_1(x,z) := K_{\boldsymbol{\lambda}}(x,z) = \sum_{k=0}^\infty \lambda_k^2\,\cos 2\pi k\,(x-z)$$

*with $\lambda_k \ge 0$, and tensor product kernels $K_{\boldsymbol{\lambda}}^d$ as before, it is sufficient for polynomial tractability in the randomized setting that the following holds:*

*(a)* $\sum_{k=0}^\infty \lambda_k^2 = 1$,

*(b)* $\sigma_{\boldsymbol{\lambda}} := \sum_{k=1}^\infty k\,\lambda_k^2 < \infty$.

*In detail, there exists a universal constant $C' > 0$ such that*

$$n^{\mathrm{ran}}(\varepsilon, \mathcal{H}(K_{\boldsymbol{\lambda}}^d) \hookrightarrow L_\infty(\mathbb{T}^d), \Lambda^{\mathrm{all}}) \le C'\,(1+\sigma_{\boldsymbol{\lambda}}^2)\,\frac{d\,(1+\log d)}{\varepsilon^2}\,.$$

*Proof.* Condition (a) is for the normalization of the initial error, see (ii) in Theorem 3.17.

We first check, when assumption (iii) of Theorem 3.17 holds with $R_0 = \frac{1}{2}$, that is, the inequality $K_{\boldsymbol{\lambda}}(x,z) \geq \exp(-\alpha\, d_{\mathbb{T}}(x,z))$ is valid for all $x, z \in \mathbb{T}$. It suffices to show that

$$h(x) := \exp(\alpha\, x)\, K_{\boldsymbol{\lambda}}(x,0)$$

is monotonously increasing for $x > 0$, noting $h(0) = 1$ by (a). Condition (b) guarantees differentiability of $K_{\boldsymbol{\lambda}}(\cdot,0)$ with absolute convergence of the resulting series of sine functions. Moreover, for the derivative of $h$ we obtain

$$\begin{aligned} h'(x) &= \exp(\alpha\, x) \left[\alpha \sum_{k=0}^{\infty} \lambda_k^2 \cos 2\pi k\, x - 2\pi \sum_{k=1}^{\infty} k\, \lambda_k^2 \sin 2\pi k\, x\right] \\ &\geq \exp(\alpha\, x) \left[\alpha \left(\lambda_0^2 - \sum_{k=1}^{\infty} \lambda_k^2\right) - 2\pi \sum_{k=1}^{\infty} k\, \lambda_k^2\right]. \end{aligned}$$

Positivity of the left-hand term in $[\ldots]$ is ensured if $\lambda_0^2 > \frac{1}{2}$ holds in addition to (a). The right-hand term in $[\ldots]$ is finite thanks to condition (b), hence we can choose

$$\alpha := \frac{2\pi\, \sigma_{\boldsymbol{\lambda}}}{2\,\lambda_0^2 - 1}$$

to guarantee non-negativity of $h'(x)$. This gives us $K_{\boldsymbol{\lambda}}(x,z) \geq \exp(-\alpha\,|x-z|)$ as intended. Restricting to the case $\lambda_0^2 \geq \frac{2}{3}$ we have a better control over the constants, in that case getting $\alpha \leq 6\pi\, \sigma_{\boldsymbol{\lambda}}$. Hence by Theorem 3.17,

$$\begin{aligned} n^{\mathrm{ran}}(\varepsilon, \mathcal{H}(K_{\boldsymbol{\lambda}}^d) \hookrightarrow L_\infty(\mathbb{T}^d), \Lambda^{\mathrm{all}}) &\leq C\,(1 + (6\,\pi)^2\, \sigma_{\boldsymbol{\lambda}}^2)\, d\,(1 + \log d)\, \varepsilon^{-2} \\ &\leq C_0\,(1 + \sigma_{\boldsymbol{\lambda}}^2)\, d\,(1 + \log d)\, \varepsilon^{-2}\,. \end{aligned}$$

If $\lambda_0^2 < \frac{2}{3}$,[23] we compare $K_{\boldsymbol{\lambda}}$ to a kernel $K_{\boldsymbol{\kappa}}$ with $\kappa_0^2 = \frac{2}{3}$ and $\kappa_k^2 = \lambda_k^2/c$ for $k \in \mathbb{N}$, where $c := 3\,(1 - \lambda_0^2) > 1$ is a scaling factor such that (a) holds for $\boldsymbol{\kappa}$ as well. Besides, (b) is inherited from $\boldsymbol{\lambda}$. This shows the existence of a constant $0 < \alpha \leq 6\pi\, \sigma_{\boldsymbol{\kappa}} \leq 6\pi\, \sigma_{\boldsymbol{\lambda}}$ such that $K_{\boldsymbol{\kappa}}(x,0) \geq \exp(-\alpha\, x)$ for $x \geq 0$. Note that

$$\begin{aligned} K_{\boldsymbol{\lambda}}(x,0) &= \lambda_0^2 + c\,\left(K_{\boldsymbol{\kappa}}(x,0) - \tfrac{2}{3}\right) \\ &\geq c\, \exp(-\alpha\, x) - \left(\tfrac{2}{3} c - \lambda_0^2\right) \\ &= 3\,(1 - \lambda_0^2)\, \exp(-\alpha\, x) - (2 - 3\lambda_0^2)\,. \end{aligned}$$

If we choose $\beta = 4\alpha$, for $x \leq R_0 := \min\left\{\alpha^{-1}\, \log\tfrac{9}{8},\, \tfrac{1}{2}\right\}$ we can finish with the estimate

$$K_{\boldsymbol{\lambda}}(x,0) \geq \exp(-\beta\, x)\,.$$

[23] In the case $\lambda_0^2 \leq \frac{1}{2}$ we need a local estimate because we can not a priori exlude negative or vanishing values for the kernel function $K(\mathbf{x},\mathbf{z})$ for far apart points $\mathbf{x}$ and $\mathbf{z}$. In the case $\frac{1}{2} < \lambda_0^2 < \frac{2}{3}$ a localized view will make better constants possible, in the end we aim for a universal constant $C'$.

Here we used that with $\beta > \alpha$ and $\gamma > 0$, for $0 \le x \le \alpha^{-1} \log\left((1-\alpha/\beta)\,(1+1/\gamma)\right)$ the following inequality holds,

$$\exp(-\beta x) \le (1+\gamma)\,\exp(-\alpha x) - \gamma\,,$$

in our case $\beta = 4\alpha$ and $\gamma = 2 - 3\lambda_0^2 \le 2$. (A proof for this inequality can be done by observing that RHS/LHS as a function in $x$ is monotonously growing for small $x > 0$.) By this, from Theorem 3.17 we obtain the complexity bound

$$\begin{aligned} n^{\mathrm{ran}}(\varepsilon, \mathcal{H}(K_{\boldsymbol{\lambda}}^d) \hookrightarrow L_\infty(\mathbb{T}^d), \Lambda^{\mathrm{all}}) &\le C\,(1+\beta^2-\log 2R_0)\,d\,(1+\log d)\,\varepsilon^{-2} \\ &\le C_1\,(1+\alpha^2+\log\alpha)\,d\,(1+\log d)\,\varepsilon^{-2} \\ &\le C_2\,(1+\sigma_{\boldsymbol{\lambda}}^2)\,d\,(1+\log d)\,\varepsilon^{-2}\,. \end{aligned}$$

Finally, the constant in the corollary is $C' := \max\{C_0, C_2\}$. □

**Example: Korobov Spaces**

We apply the above results to unweighted Korobov spaces. In the framework of this section, these are spaces $\mathcal{H}_r^{\mathrm{Kor}}(\mathbb{T}^d) := \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d)$ with $\lambda_0 = \sqrt{\beta_0}$ and $\lambda_k = \sqrt{\beta_1}\,k^{-r}$ for $k \in \mathbb{N}$, where $\beta_0, \beta_1 > 0$. For integers $r \in \mathbb{N}$, the Korobov space norm can be given in a natural way in terms of weak partial derivatives (instead of Fourier coefficients), in the one-dimensional case we have

$$\|f\|^2_{\mathcal{H}_r^{\mathrm{Kor}}(\mathbb{T})} = \beta_0^{-1}\left|\int_{\mathbb{T}} f(x)\,\mathrm{d}x\right|^2 + \beta_1^{-1}\,(2\pi)^{-2r}\,\|f^{(r)}\|_2^2\,.$$

The $d$-dimensional case is a bit more complicated, in a squeezed way, the norm is

$$\|f\|^2_{\mathcal{H}_r^{\mathrm{Kor}}(\mathbb{T}^d)} = \sum_{J\subseteq[d]} \beta_0^{-(d-\#J)}\,(\beta_1^{-1}\,(2\pi)^{-2r})^{\#J} \left\|\int_{\mathbb{T}^{[d]\setminus J}} \Big(\prod_{j\in J}\partial_j^r\Big) f(\mathbf{x})\;\mathrm{d}\mathbf{x}_{[d]\setminus J}\right\|^2_{L_2(\mathbb{T}^J)},$$

see Novak and Woźniakowski [55, Sec A.1] for details on the derivation of this representation of the norm. There one can also find some information on the historical background concerning these spaces. It should be pointed out that in the same book tractability for $L_2$-approximation of Korobov functions based on $\Lambda^{\mathrm{all}}$ has been studied [55, pp. 191–193], in that case randomization does not help a lot.

The condition $r > \frac{1}{2}$ is necessary and sufficient for the existence of a reproducing kernel (and the embedding $\mathcal{H}_r^{\mathrm{Kor}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)$ to be compact), then

$$\sum_{k=1}^{\infty} \lambda_k^2 = \beta_1 \sum_{k=1}^{\infty} k^{-2r} = \beta_1\,\zeta(2r)$$

with the Riemann zeta function $\zeta$. Assuming

$$\beta_0 + \beta_1\,\zeta(2r) = 1\,, \tag{3.4.4}$$

the initial error will be constant 1 in all dimensions. Furthermore, with $\beta_1 > 0$ we have the curse of dimensionality for the deterministic setting, see Theorem 3.15.

**Theorem 3.19.** *Consider unweighted Korobov spaces $\mathcal{H}_r^{\mathrm{Kor}}(\mathbb{T}^d) = \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d)$ as described above. For smoothness $r > \frac{1}{2}$, fixing $\beta_0, \beta_1 \geq 0$ such that the initial error is constant 1 for all dimensions, we have polynomial tractability for the uniform approximation with Monte Carlo, in detail,*

$$n^{\mathrm{ran}}(\varepsilon, \mathcal{H}_r^{\mathrm{Kor}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d), \Lambda^{\mathrm{all}}) \preceq \begin{cases} d\,(1+\log d)\,\varepsilon^{-2} & \text{for } r > 1, \\ d\,(1+\log d)^3\,\varepsilon^{-2}\,(1+\log \varepsilon^{-1})^2 & \text{for } r = 1, \\ d^{1/(r-1/2)-1}\,(1+\log d)\,\varepsilon^{-1/(r-1/2)} & \text{for } \frac{1}{2} < r < 1. \end{cases}$$

*The hidden constant may depend on $r$.*

*Proof.* We start with the easiest case $r > 1$. By (3.4.4) we have $\beta_1 \leq 1/\zeta(2r-1) < 1$, thus we satisfy (b) in Corollary 3.18 with

$$\sigma_{\boldsymbol{\lambda}} = \sum_{k=1}^{\infty} k\,\lambda_k^2 = \beta_1 \sum_{k=1}^{\infty} k^{-(2r-1)} \leq \zeta(2r-1)\,,$$

and obtain (with the constant $C'$ from the corollary)

$$n^{\mathrm{ran}}(\varepsilon, \mathcal{H}_r^{\mathrm{Kor}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d), \Lambda^{\mathrm{all}}) \leq C'\,(1+\zeta(2r-1)^2)\,\frac{d\,(1+\log d)}{\varepsilon^2}\,.$$

For $\frac{1}{2} < r \leq 1$, the quantity $\sigma_{\boldsymbol{\lambda}}$ is infinite. Therefore we apply the fundamental Monte Carlo method from Proposition 3.1 to a finite dimensional subspace of finite Fourier sums up to frequencies $k_j \leq m$ in each dimension. With the orthonormal basis $\{\psi_{\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$ of $\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d)$, see (3.4.1), for $f \in \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d)$ we define

$$f_m := \sum_{\substack{\mathbf{k}\in\mathbb{Z}^d \\ |\mathbf{k}|_\infty \leq m}} \langle \psi_{\mathbf{k}}, f\rangle_{\mathcal{H}_{\boldsymbol{\lambda}}}\,\psi_{\mathbf{k}}\,.$$

Taking a Monte Carlo method $(A_n^\omega)_{\omega\in\Omega}$ with $A_n^\omega(f) = A_n^\omega(f_m)$, we can estimate the error for $f \in \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d)$ by

$$e((A_n^\omega)_\omega, f) \leq \|f - f_m\|_\infty + e((A_n^\omega)_\omega, f_m)\,.$$

For this term to be bounded from above by $\varepsilon$, we desire both summands to be bounded from above by $\varepsilon/2$.

By the worst case error formula (3.3.4), together with the kernel representation (3.4.2), for $\|f\|_{\mathcal{H}_{\boldsymbol{\lambda}}} \leq 1$ it easily follows

$$\|f - f_m\|_\infty^2 \;\leq\; 1 - \sum_{\substack{\mathbf{k}\in\mathbb{Z}^d \\ |\mathbf{k}|_\infty \leq m}} \lambda_{\mathbf{k}}^2 \;=\; 1 - \left(\sum_{k=0}^{m} \lambda_k^2\right)^d\,.$$

In our particular situation with $\lambda_k^2 = \beta_1\,k^{-2r}$ for $k \in \mathbb{N}$, we can estimate

$$\sum_{k=m+1}^{\infty} \lambda_k^2 \leq \beta_1 \int_m^\infty t^{-2s}\,\mathrm{d}t = \frac{\beta_1}{2r-1}\,m^{-(2r-1)}\,.$$

Hence, together with $\sum_{k=0}^{\infty} \lambda_k^2 = 1$, we obtain

$$\begin{aligned} \|f - f_m\|_{\mathcal{H}_{\boldsymbol{\lambda}}}^2 &\leq 1 - \left(1 - \frac{\beta_1}{2r-1}\, m^{-(2r-1)}\right)^d \\ \left[\text{for } \tfrac{\beta_1}{2r-1}\, m^{-(2r-1)} < \tfrac{1}{2}\right] \quad &\overset{(*)}{\leq} 1 - \exp\left(-\log 2\, \frac{\beta_1}{r-1/2}\, d\, m^{-(2r-1)}\right) \\ &\leq \log 2\, \frac{\beta_1}{r-1/2}\, d\, m^{-(2r-1)}\,. \end{aligned}$$

Here we used $1 - t > \exp(-2\,(\log 2)\, t)$ for $x \in (0, \frac{1}{2})$. Choosing

$$m := \left\lceil \left(4\,(\log 2)\, \frac{\beta_1}{r-1/2}\, d\, \varepsilon^{-1}\right)^{1/(2r-1)} \right\rceil\,, \tag{3.4.5}$$

step $(*)$ is actually valid, and we bound $\|f - f_m\|_\infty \leq \varepsilon/2$.

For the error analysis of $e((A_n^\omega)_\omega, f_m)$, we need to understand the restricted approximation problem

$$\mathrm{APP} : \mathcal{H}_{\boldsymbol{\lambda}'}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)\,, \quad \text{where } \lambda_k' := \lambda_k\, \mathbb{1}[k \leq m].$$

The method $A_n$ shall be the fundamental Monte Carlo approximation method from Proposition 3.1 applied to this problem. The initial error is smaller than 1, so we cannot apply Corollary 3.18 directly to this problem. Therefore, consider another space $\mathcal{H}_{\boldsymbol{\kappa}}(\mathbb{T}^d)$ with $\kappa_k = \lambda_k'$ for $k \in \mathbb{N}$ and $\kappa_0 := \sqrt{1 - \sum_{k=1}^m \lambda_k^2} > \lambda_0$. The initial error of the approximation problem

$$\mathrm{APP} : \mathcal{H}_{\boldsymbol{\kappa}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)$$

is then properly normalized by construction. Applying Proposition 3.1 to $\mathcal{H}_{\boldsymbol{\lambda}'}(\mathbb{T}^d)$ and $\mathcal{H}_{\boldsymbol{\kappa}}(\mathbb{T}^d)$ means determining the expected $L_\infty$-norm of the corresponding Gaussian processes $\Psi^{(\boldsymbol{\lambda}',d)}$ and $\Psi^{(\boldsymbol{\kappa},d)}$, respectively. For better comparison it is useful to represent these via the Fourier basis $\{\varphi_{\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$ of $L_2(\mathbb{T}^d)$,

$$\Psi^{(\boldsymbol{\lambda}',d)} = \sum_{\substack{\mathbf{k}\in\mathbb{Z}^d \\ |\mathbf{k}|_\infty \leq m}} 2^{-|\mathbf{k}|_0/2}\, \lambda_{\mathbf{k}}\, X_{\mathbf{k}}\, \varphi_{\mathbf{k}}\,, \qquad \text{and} \qquad \Psi^{(\boldsymbol{\kappa},d)} = \sum_{\substack{\mathbf{k}\in\mathbb{Z}^d \\ |\mathbf{k}|_\infty \leq m}} 2^{-|\mathbf{k}|_0/2}\, \kappa_{\mathbf{k}}\, X_{\mathbf{k}}\, \varphi_{\mathbf{k}}\,,$$

where the $X_{\mathbf{k}}$ are iid standard Gaussian random variables. Note that, by construction, $\kappa_{\mathbf{k}} \geq \lambda_{\mathbf{k}}'$ for $k \in \mathbb{Z}^d$, so we have $\mathbb{E}\,\|\Psi^{(\boldsymbol{\lambda}',d)}\|_\infty \leq \mathbb{E}\,\|\Psi^{(\boldsymbol{\kappa},d)}\|_\infty$, see Lemma A.2. Consequently, complexity bounds from applying Corollary 3.18 to $\mathcal{H}_{\boldsymbol{\kappa}}(\mathbb{T}^d)$ also hold for $\mathcal{H}_{\boldsymbol{\lambda}'}(\mathbb{T}^d)$. This gives

$$\begin{aligned} n^{\mathrm{ran}}(\varepsilon, \mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)) &\leq n^{\mathrm{ran}}(\varepsilon/2, \mathcal{H}_{\boldsymbol{\lambda}'}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)) \\ &\leq n^{\mathrm{ran}}(\varepsilon/2, \mathcal{H}_{\boldsymbol{\kappa}}(\mathbb{T}^d) \hookrightarrow L_\infty(\mathbb{T}^d)) \\ &\leq 4\, C'\, (1 + \sigma_{\boldsymbol{\kappa}}^2)\, \frac{d\,(1 + \log d)}{\varepsilon^2}\,. \end{aligned} \tag{3.4.6}$$

It remains to estimate $\sigma_{\boldsymbol{\kappa}}$:

$$\begin{aligned}
\sigma_{\boldsymbol{\kappa}} &= \sum_{k=1}^{\infty} k\, \lambda_k'^2 \\
&= \beta_1 \sum_{k=1}^{m} k^{-(2r-1)} \\
[\text{neglect } \beta_1 < 1] \qquad &\le 1 + \int_1^m t^{-(2r-1)}\, \mathrm{d}t \\
&\le \begin{cases} 1 + \log m & \text{for } r = 1, \\ \frac{1}{2(1-r)}\, m^{2(1-r)} & \text{for } \frac{1}{2} < r < 1. \end{cases}
\end{aligned}$$

By the choice of $m$, see (3.4.5), putting this into (3.4.6), we obtain the final upper bound with constants that depend on $r$. □

**Remark 3.20** (Loss of smoothness)**.** In the case $r > 1$, with the simple approach from Section 3.3.1 we loose smoothness $\frac{1}{2}$. This stresses the non-interpolatory nature of the method. In detail, check that the Gaussian process $\Psi$ associated to $\mathcal{H}_r^{\mathrm{Kor}}$ (for any equivalent norm) lies almost surely in $\mathcal{H}_s^{\mathrm{Kor}}$ for $r - s > \frac{1}{2}$, and it is almost surely not in $\mathcal{H}_s^{\mathrm{Kor}}$ for $r - s \le \frac{1}{2}$. The argument is similar to that in Remark 3.2.

**Remark 3.21** (On the case of smaller smoothness)**.** We had some difficulties with the case of smaller smoothness $\frac{1}{2} < r \le 1$. However, there is some indication that nevertheless the associated Gaussian process $\Psi$ is bounded. Then we could apply the simple approach from Section 3.3.1, and that way obtain better complexity bounds than in Theorem 3.19.

Sufficient and necessary conditions on $\boldsymbol{\lambda}$ for boundedness of the univariate Gaussian Fourier series $\Psi$ associated to $\mathcal{H}_{\boldsymbol{\lambda}}(\mathbb{T})$ are known, see Adler [1, Thm 1.5]. The book of Marcus and Pisier [45] contains a lot more information on random Fourier series and could serve as a starting point for further research.

There is a second hint. Plots of the one-dimensional kernel $K_1$ for $\mathcal{H}_r^{\mathrm{Kor}}(\mathbb{T})$ nourish the conjecture that for $r > \frac{1}{2}$ we can find an estimate

*(iii)'* $K_1(x,z) \ge \exp(-\alpha\, d_{\mathbb{T}}(x,z)^p)$ *for* $x, z \in \mathbb{T}$ *with* $d_{\mathbb{T}}(x,z) \le R_0$,

with $0 < p < 2r - 1$ and $\alpha > 0$. An adapted version of Theorem 3.17 could give the desired upper bounds.

Anyway, we were mainly interested in showing the superiority of Monte Carlo approximation over deterministic approximation in terms of tractability, and we have been successful for the whole range of continuous Korobov functions.

**Remark 3.22** (Combined methods)**.** The upper bounds in Theorem 3.19 do not give the optimal order of convergence for the approximation of Korobov functions. The rate of convergence we can guarantee by this is only $e^{\mathrm{ran}}(n) \preceq n^{-1/2}$ for $r > 1$, and it can be arbitrarily bad for low smoothness $r$ close to $\frac{1}{2}$. In fact, similar to $L_q$ approximation in Fang and Duan [15], $1 < q < \infty$, one will obtain a rate

$(n^{-1}(\log n)^{d-1})^r \preceq e^{\mathrm{ran}}(n) \preceq (n^{-1}(\log n)^{d-1})^r \sqrt{\log n}$.[24] Proofs in that direction still can be performed using Maiorov's discretization technique, but more direct methods will give better constants with simpler proofs.

In detail, propose an explicit Monte Carlo approximation method $A_n$ with the following properties:

(1) The most relevant Fourier coefficients, belonging to the indices $I_{\mathrm{det}} \subset \mathbb{Z}^d$, are approximated exactly at the cost of evaluating $n_{\mathrm{det}} = \#I_{\mathrm{det}}$ functionals.

(2) Fourier coefficients of medium importance, belonging to $I_{\mathrm{ran}} \subset \mathbb{Z}^d$ are approximated altogether now using the fundamental Monte Carlo approximation method from Proposition 3.1 with $n_{\mathrm{ran}} \ll \#I_{\mathrm{ran}}$ random functionals, the total information cost is $n = n_{\mathrm{det}} + n_{\mathrm{ran}}$.

(3) The remaining Fourier coefficients, for indices $\mathbb{Z}^d \setminus (I_{\mathrm{det}} \cup I_{\mathrm{ran}})$, are ignored.

The effect of truncation (3) should be estimated with methods from Section 3.3.4, see also Cobos et al. [12] or Kuo et al. [40]. The Monte Carlo part (2) could be treated with methods on Gaussian fields, see Section 3.3.3 and Adler [1], or maybe Marcus and Pisier [45]. Note that by the deterministic part (1) we will likely have no tensor product structure for the analysis of the Monte Carlo part (2).

### 3.4.3 Final Remarks on the Initial Error

For unweighted tensor product problems as in the above two subsections, the assumption of a normalized initial error is crucial for the new approach to work. If not, that is, if

$$e(0, \mathcal{H}(K_1) \hookrightarrow L_\infty(D_1)) = \sup_{x \in D_1} \sqrt{K_1(x,x)} =: 1 + \gamma > 1\,,$$

we have

$$e(0, \mathcal{H}(K_d) \hookrightarrow L_\infty(D_d)) = (1+\gamma)^d\,,$$

where $K_d$ is the product kernel and $D_d := \bigtimes_{j=1}^d D_1$. Then for the Gaussian field $\Psi$ with covariance function $K_d$ we have

$$\mathbb{E}\,\|\Psi\|_\infty \geq \sup_{\mathbf{x} \in D_d} \mathbb{E}\,|\Psi_{\mathbf{x}}| = \sqrt{\frac{2}{\pi}}\,(1+\gamma)^d\,.$$

In this situation Lemma 3.3 can only give an impractical upper complexity bound that grows exponentially in $d$ for fixed $\varepsilon > 0$. However, if the constant function $f = 1$ is normalized in $\mathcal{H}(K_1)$, but $\mathcal{H}(K_1)$ is non-trivial and contains more than just constant functions, then $\sup_{x \in D_1} \sqrt{K_1(x,x)} > 1$, contrary to our requirements.

Therefore we must accept that the constant function $f = 1$ cannot be a normalized function in $\mathcal{H}(K_d)$ if we want to break the curse for the $L_\infty$-approximation with the present tools. This stands in contrast to many other problems:

[24] This topic is part of a cooperation with Glenn Byrenheid and Van Kien Nguyen, soon to be published.

- Take the $L_2$ approximation of periodic Korobov spaces, see e.g. Novak and Woźniakowski [55, pp. 191–193]. The initial error is 1 iff the largest singular value is 1, so it is natural to let the constant function $f = 1$ be normalized within the input space $\mathcal{H}$, hence it lies on the boundary of the input set.

- Consider multivariate integration over Korobov spaces, see for example Novak and Woźniakowski [57, Chap 16]. The integral is actually the Fourier coefficient belonging to the constant function. The initial error is properly normalized iff the constant function $f = 1$ has norm 1.

I wish to thank my colleagues David Krieg, and Van Kien Nguyen (meanwhile Dr. rer. nat.), for making me aware of the difference to these particular two situations.

# Chapter 4

# $L_1$-Approximation of Monotone Functions

For the $L_1$-approximation of $d$-variate monotone functions (and also monotone Boolean functions) by function values, the *curse of dimensionality* holds in the deterministic setting, see Hinrichs, Novak, and Woźniakowski [28], and Section 4.2.2. For the randomized setting we still have *intractability*, i.e. the problem is *not weakly tractable*, see Section 4.3 where we improve known lower bounds, the new bounds now exhibit a meaningful $\varepsilon$-dependency. Yet randomization may reduce the complexity significantly, for any fixed tolerance $\varepsilon$ the complexity depends exponentially on $\sqrt{d}$, roughly, see Section 4.4 for the analysis of a known algorithm for Boolean functions, and a new extension to real-valued monotone functions that is based on a Haar wavelet decomposition.

## 4.1 The Setting and Background

Within this chapter we mainly consider the $L_1$-approximation of $d$-variate *monotone* functions using function values $\Lambda^{\mathrm{std}}$ as information,[1]

$$\mathrm{APP} : F_{\mathrm{mon}}^d \hookrightarrow L_1([0,1]^d)\,,$$

where the input set

$$F_{\mathrm{mon}}^d := \{f : [0,1]^d \to [0,1] \mid \mathbf{x} \le \tilde{\mathbf{x}} \Rightarrow f(\mathbf{x}) \le f(\tilde{\mathbf{x}})\}$$

consists of monotonously increasing functions with respect to the partial order on the domain. For $\mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^d$, the partial order is defined by

$$\mathbf{x} \le \tilde{\mathbf{x}} :\Leftrightarrow x_j \le \tilde{x}_j \text{ for all } j = 1, \dots, d\,. \tag{4.1.1}$$

[1] The input set contains different functions that belong to the same equivalence class in $L_1[0,1]^d$. Actually, we do not care about these equivalence classes but only need the $L_1$-norm as a seminorm. Furthermore, function evaluations are discontinuous functionals, but monotonicity provides a regularization to this type of information in some other useful ways such that deterministic approximation is actually possible.

This problem is closely related to the approximation of *Boolean* monotone functions

$$F_{\circ|}^d := \{f : \{0,1\}^d \to \{0,1\} \mid \mathbf{x} \le \tilde{\mathbf{x}} \Rightarrow f(\mathbf{x}) \le f(\tilde{\mathbf{x}})\}\,.$$

One can identify $F_{\circ|}^d$ with a subclass $\widetilde{F}_{\circ|}^d \subseteq F_{\mathrm{mon}}^d$ if we split $[0,1]^d$ into $2^d$ subcubes indexed by $\mathbf{i} \in \{0,1\}^d$,

$$C_{\mathbf{i}} := \bigtimes_{j=1}^{d} I_{i_j}\,, \quad \text{where } I_0 := [0, \tfrac{1}{2}) \text{ and } I_1 := [\tfrac{1}{2}, 1]. \tag{4.1.2}$$

Then, for any Boolean function $f : \{0,1\}^d \to \{0,1\}$, we obtain a subcubewise constant function $\tilde{f} : [0,1]^d \to [0,1]$ by setting $\tilde{f}|_{C_{\mathbf{i}}} := f(\mathbf{i})$. If $f$ is monotone then so is $\tilde{f}$. The corresponding distance between two Boolean functions $f_1, f_2 : \{0,1\}^d \to \{0,1\}$ is

$$\mathrm{dist}(f_1, f_2) := \frac{1}{2^d}\,\#\{\mathbf{i} \in \{0,1\}^d \mid f_1(\mathbf{i}) \neq f_2(\mathbf{i})\}\,. \tag{4.1.3}$$

The metric space of Boolean functions shall be named $G_{\circ|}^d$, we consider the approximation problem

$$\mathrm{APP} : F_{\circ|}^d \hookrightarrow G_{\circ|}^d\,.$$

Note that the metric on $G_{\circ|}^d$ corresponds to the $L_1$-distance of the associated subcubewise constant functions defined on $[0,1]^d$. Another way to think of the metric on $G_{\circ|}^d$ is as the induced metric for $G_{\circ|}^d$ as a subset of the Banach space $L_1(\mathrm{unif}\{0,1\}^d)$.[2]

Approximation of monotone functions is not a linear problem becasue the set $F_{\mathrm{mon}}^d$ is not symmetric: For non-constant functions $f \in F_{\mathrm{mon}}^d$, the negative $-f$ is not contained in $F_{\mathrm{mon}}^d$ as it will be monotonously decreasing. The monotonicity assumption is very different from common smoothness assumptions, yet it implies many other nice properties, see for example Alberti and Ambrosio [3]. Integration and Approximation of monotone functions has been studied in several papers [28, 53, 62]. Monotonicity can also be an assumption for statistical problems [18, 69]. Similarly, a structural assumption could be convexity (more generally: $k$-monotonicity), numerical problems with such properties have been studied for example in [11, 28, 32, 36, 54].

Within this research, Boolean monotone functions are considered in order to obtain lower bounds for the Monte Carlo approximation of real-valued monotone functions, see Section 4.3. We will show that the approximation of monotone (Boolean) functions is *not weakly tractable*, i.e. *intractable* in the IBC sense.[3] General Boolean functions $f : \{0,1\}^d \to \{0,1\}$ are of interest for logical networks and cryptographic applications. Monotone Boolean functions in particular constitute a widely studied topic in computer science and discrete mathematics with connections to graph theory amongst others. Much research has been done on different effective ways of exact representation of Boolean functions, see the survey paper of Korshunov [37]. On

[2] This property is the reason why it is convenient to consider real-valued monotone functions with range $[0,1]$. It is only in Section 4.4 that we switch to the range $[-1,+1]$ because there we use linear approximation methods.

[3] In learning theory there exist many similar sounding notions like *weak learnability*, which, however, have different meanings.

the other hand, the approximation of monotone Boolean functions (or subclasses thereof) is a good example for learning theory [4, 8, 9, 33, 59]. In cryptography one is interested in finding classes of easily representable Boolean functions (not only monotone functions) that are hard to learn from examples [34, 67], so the motivation is different from the motivation for tractability studies in IBC. Different perspectives on similar problems explain the sometimes unfamiliar way of presenting results in other scientific communities, see for instance Section 4.3.

Finally, we give some examples of naturally occurring monotone functions.

**Example 4.1** (Application of monotone functions)**.** Think of a complex technical system with $d$ components. Some of these components could be damaged but the system as a whole would still work. However, if more critical components fail, the system will fail as well. It is a natural assumption that, once the system stopped working, it will not come back to life after more components break – this, in fact, is monotonicity.

We are interested in predicting when the machine will cease to function. Our framework fits to a test environment where we can manually deactivate components and check whether the system still does its job. We are interested in a good (randomized) strategy to test the system. This approach could work for the testing of uncritical applications where a test environment can be set up.

For rather critical systems such as aircrafts or running systems, we need to learn from bad experience (that we actually wish to avoid). Different components will have different probabilities of failing, and it is with these probabilities that we obtain samples from which we can learn. On the other hand, not every case is equally important, so we judge the approximation of the Boolean function by the probability of an event occurring where the prediction fails. This situation fits better to the framework that we have for example in Bshouty and Tamon [9]. (From that paper we know the method presented in Section 4.4.1.)

This picture can be extended to real-valued monotone functions $f : [0,1]^d \to [0,1]$. Now, each of the $d$ components of a system can work at a different level, we are interested in how much this affects the performance of the entire system. When thinking of a home computer, the question could be how much the PC slows down in different situations.

## 4.2 First Simple Estimates

In view of the order of convergence for the approximation of real-valued monotone functions, see Section 4.2.1, randomization does not help. Actually, for small errors $\varepsilon > 0$ the very simple deterministic algorithm to be found in this section is the best method we know, the randomized methods from Section 4.4.2 will only help for larger $\varepsilon$.

In Section 4.2.2 we cite a result of Hinrichs, Novak, and Woźniakowski [28], which states that the approximation of real-valued monotone functions suffers from the curse of dimensionality in the deterministic setting. A similar statement holds for the approximation of Boolean monotone functions as well.

### 4.2.1 The Classical Approach – Order of Convergence

The integration problem for monotone functions,

$$\mathrm{INT} : F_{\mathrm{mon}}^d \to \mathbb{R}, \quad f \mapsto \int_{[0,1]^d} f \,\mathrm{d}\mathbf{x}\,,$$

based on standard information $\Lambda^{\mathrm{std}}$, is an interesting numerical problem, where in the randomized setting adaption makes a difference for the order of convergence (at least for $d = 1$), but non-adaptive randomization helps only for $d \geq 2$ to improve the convergence compared to deterministic methods. In the univariate case Novak [53] showed

$$e^{\mathrm{ran,ada}}(n, \mathrm{INT}, F_{\mathrm{mon}}^1) \asymp n^{-3/2} \\ \prec e^{\mathrm{ran,nonada}}(n, \mathrm{INT}, F_{\mathrm{mon}}^1) \asymp e^{\mathrm{det}}(n, \mathrm{INT}, F_{\mathrm{mon}}^1) \asymp n^{-1}\,.$$

Papageorgiou [62] examined the integration for $d$-variate monotone functions, for dimensions $d \geq 2$ we have

$$e^{\mathrm{ran,ada}}(n, \mathrm{INT}, F_{\mathrm{mon}}^d) \asymp n^{-\frac{1}{d}-\frac{1}{2}} \\ \preceq e^{\mathrm{ran,nonada}}(n, \mathrm{INT}, F_{\mathrm{mon}}^d) \preceq n^{-\frac{1}{2d}-\frac{1}{2}} \\ \prec e^{\mathrm{det}}(n, \mathrm{INT}, F_{\mathrm{mon}}^d) \asymp n^{-\frac{1}{d}}\,,$$

where the hidden constants depend on $d$. It is an open problem to find lower bounds for the non-adaptive Monte Carlo error that actually show that adaption is better for $d \geq 2$ as well, but from the one-dimensional case we conjecture it to be like that.

For the $L_1$-approximation, the order of convergence does not reveal any differences between the various algorithmic settings. Applying Papageorgiou's proof technique to the problem of $L_1$-approximation, we obtain the following theorem.

**Theorem 4.2.** *For the $L_1$-approximation of monotone functions, for fixed dimension $d$ and $n \to \infty$, we have the following asymptotic behaviour,*

$$e^{\mathrm{ran}}(n, \mathrm{APP}, F_{\mathrm{mon}}^d) \asymp e^{\mathrm{det}}(n, \mathrm{APP}, F_{\mathrm{mon}}^d) \asymp n^{-\frac{1}{d}}\,.$$

*This holds also for varying cardinality.*

*Proof.* We split $[0,1]^d$ into $m^d$ subcubes indexed by $\mathbf{i} \in \{0, 1, \ldots, m-1\}^d$:

$$C_{\mathbf{i}} := \bigtimes_{j=1}^{d} I_{i_j}$$

where $I_i := [\frac{i}{m}, \frac{i+1}{m})$ for $i = 0, 1, \ldots, m-2$ and $I_{m-1} := [\frac{m-1}{m}, 1]$.

For the lower bounds, we consider fooling functions that are constant on each of the subcubes, in detail,

$$f|_{C_{\mathbf{i}}} = \frac{|\mathbf{i}|_1 + \delta_{\mathbf{i}}}{d(m-1)+1}$$

with $\delta_{\mathbf{i}} \in \{0,1\}$ and $|\mathbf{i}|_1 := i_1 + \ldots + i_d$. Obviously, such functions are monotonously increasing. For a deterministic algorithm using $n < m^d$ function values, when applied to such a function, we do not know the function on at least $m^d - n$ subcubes. There exist a maximal function $f_+$ and a minimal function $f_-$ fitting to the computed function values, and the diameter of information is at least

$$\|f_+ - f_-\|_1 \geq \left(1 - \frac{n}{m^d}\right) \frac{1}{d(m-1)+1}\,.$$

Hence we have the error bound

$$e^{\mathrm{det}}(n, \mathrm{APP}, F^d_{\mathrm{mon}}) \geq \frac{1}{2}\,\left(1 - \frac{n}{m^d}\right) \frac{1}{d(m-1)+1}\,.$$

For Monte Carlo lower bounds we switch to the average case setting, the measure $\mu$ may be described by such functions where the $\delta_{\mathbf{i}}$ are independent Bernoulli variables with $\mu\{\delta_{\mathbf{i}} = 0\} = \mu\{\delta_{\mathbf{i}} = 1\} = \frac{1}{2}$. For any information $\mathbf{y}$, let $I^{\mathbf{y}} \subset \{0, \ldots, m-1\}^d$ be the set of indices $\mathbf{i}$ where we do not know anything about the function on the corresponding subcube $C_{\mathbf{i}}$. Again, $\#I^{\mathbf{y}} \geq m^d - n$, and for any output $g : [0,1]^d \to [0,1]$ we have the following estimate on the local average error with respect to the conditional distribution $\mu_{\mathbf{y}}$:

$$\begin{aligned}
&\int \|f - g\|_1 \, \mu_{\mathbf{y}}(\mathrm{d}f) \\
&\qquad \geq \sum_{\mathbf{i} \in I^{\mathbf{y}}} \frac{1}{2} \int_{C_{\mathbf{i}}} \underbrace{\left( \left| \frac{|\mathbf{i}|_1}{d(m-1)+1} - g(\mathbf{x}) \right| + \left| \frac{|\mathbf{i}|_1 + 1}{d(m-1)+1} - g(\mathbf{x}) \right| \right)}_{\geq \frac{1}{d\,(m-1)+1}} \mathrm{d}\mathbf{x} \\
&\qquad\qquad \geq \frac{1}{2}\,\left(1 - \frac{n}{m^d}\right) \frac{1}{d(m-1)+1}\,.
\end{aligned}$$

Averaging over the information $\mathbf{y}$, we obtain the same lower bound in this average setting as in the worst case setting, by virtue of Proposition 1.2 (Bakhvalov's technique), this is a lower bound for the Monte Carlo error. Note that this lower bound is an estimate for the conditional error by a convex function $\hat{\varepsilon}(\bar{n}) := c_{m,d}\,(1 - \bar{n}/m^d)_+$, notably it holds for methods with varying cardinality $n(\mathbf{y})$, putting $\bar{n} = n(\mathbf{y})$. Hence by Lemma 1.6 we reason the alike error bounds for Monte Carlo methods with varying cardinality.
Choosing $m := \lceil \sqrt[d]{2n} \rceil$, we obtain the general lower bound

$$e^{\mathrm{det}}(n, \mathrm{APP}, F^d_{\mathrm{mon}}) \geq e^{\mathrm{ran}}(n, \mathrm{APP}, F^d_{\mathrm{mon}}) \geq \frac{1}{4\,(d\,\sqrt[d]{2n} + 1)} \geq \frac{1}{12\,d}\,n^{-\frac{1}{d}}\,.$$

For the upper bounds, we give a deterministic, non-adaptive algorithm with cardinality $m^d$, i.e. when allowed to use $n$ function values, we choose $m := \lfloor \sqrt[d]{n} \rfloor$. We split the domain into $(m+1)^d$ subcubes as above, and define the output $g := A^d_m(f)$ by

$$g|_{C_{\mathbf{i}}} := \frac{1}{2}\,\left[ f\left(\frac{\mathbf{i}}{m+1}\right) + f\left(\frac{\mathbf{i}+\mathbf{1}}{m+1}\right) \right]\,,$$

here $\mathbf{1} := (1, \dots, 1) \in \mathbb{R}^d$. Without loss of generality, we assume that on the boundary of the domain we have

$$f|_{[0,1)^d \setminus (0,1)^d} = 0 \quad \text{and} \quad f|_{[0,1]^d \setminus [0,1)^d} = 1\,.$$

This means that we only need to compute $m^d$ function values on a grid in the interior $(0,1)^d$ of the domain. For each subcube we take the medium possible value based on our knowledge on the function $f$ in the lower and upper corners of that particular subcube. When analysing this algorithm, we group the subcubes into diagonals collected by index sets

$$D_{\mathbf{j}} := \{\mathbf{j} + k\,\mathbf{1} \in \{0, \dots, m\}^d \mid k \in \mathbb{Z}\}\,.$$

There are $(m+1)^d - m^d \le d\,(m+1)^{d-1}$ such diagonals, each of them can be tagged by exactly one index $\mathbf{j} \in \{0, \dots, m\}^d \setminus \{1, \dots, m\}^d$. By monotonicity we have the following estimate for the error:

$$\begin{aligned} e(A_m^d, f) &= \|f - g\|_1 \\ &\le \frac{1}{(m+1)^d} \sum_{\mathbf{j} \in \{0,\dots,m\}^d \setminus \{1,\dots,m\}^d} \underbrace{\sum_{\mathbf{i} \in D_{\mathbf{j}}} \frac{1}{2} \left[ f\left(\frac{\mathbf{i}+\mathbf{1}}{m+1}\right) - f\left(\frac{\mathbf{i}}{m+1}\right) \right]}_{=\frac{1}{2}} \\ &\le \frac{d}{2(m+1)} = \frac{d}{2(\lfloor \sqrt[d]{n} \rfloor + 1)} \le \frac{d}{2}\, n^{-\frac{1}{d}}\,. \end{aligned}$$

□

**Remark 4.3** (On the impracticality of these results)**.** The above proof yields the explicit estimate

$$\frac{1}{12\,d}\, n^{-\frac{1}{d}} \le e^{\mathrm{ran}}(n, \mathrm{APP}, F_{\mathrm{mon}}^d) \le e^{\mathrm{det}}(n, \mathrm{APP}, F_{\mathrm{mon}}^d) \le \frac{d}{2}\, n^{-\frac{1}{d}}\,.$$

At first glance, this estimate looks quite nice, with constants differing only polynomially in $d$. This optimistic view, however, collapses dramatically when switching to the notion of $\varepsilon$-complexity for $0 < \varepsilon < \frac{1}{2}$:

$$\left(\frac{1}{12\,d}\right)^d \varepsilon^{-d} \le n^{\mathrm{ran}}(\varepsilon, \mathrm{APP}, F_{\mathrm{mon}}^d) \le n^{\mathrm{det}}(\varepsilon, \mathrm{APP}, F_{\mathrm{mon}}^d) \le \left(\frac{d}{2}\right)^d \varepsilon^{-d}\,.$$

Here, the constants differ superexponentially in $d$. Of course, lower bounds for low dimensions also hold for higher dimensions, so given the dimension $d_0$, one can optimize over $d = 1, \dots, d_0$. Still, the upper bound is impractical for high dimensions since it is based on algorithms that use exponentially (in $d$) many function values.

In fact, for the deterministic setting we cannot avoid a bad $d$-dependency, as the improved lower bounds of Section 4.2.2 below show. For the randomized setting, however, we can significantly reduce the $d$-dependency (which is still high), at least as long as $\varepsilon$ is fixed,[4] see Section 4.4.2. To summarize, if we only consider the order of convergence, we might think that randomization does not help, but for high dimensions randomization actually *does* help.

[4]For small $\varepsilon \preceq 1/\sqrt{d}$ however, the best known method is the deterministic method from Theorem 4.2 above, see Remark 4.23.

### 4.2.2 Curse of Dimensionality in the Deterministic Setting

Hinrichs, Novak, and Woźniakowski [28] have shown that the integration (and hence also the $L_p$-approximation, $1 \le p \le \infty$) of monotone functions suffers from the curse of dimensionality in the deterministic setting. We want to recap their result for our particular situation.

**Theorem 4.4** (Hinrichs, Novak, Woźniakowski 2011)**.** *The $L_1$-approximation of monotone functions suffers from the curse of dimensionality in the worst case setting. In detail,*

$$e^{\mathrm{det}}(n, \mathrm{APP}, F^d_{\mathrm{mon}}, \Lambda^{\mathrm{std}}) \ge \tfrac{1}{2}\left(1 - n\,2^{-d}\right),$$

*so for $0 < \varepsilon < \frac{1}{4}$ we have*

$$n^{\mathrm{det}}(\varepsilon, \mathrm{APP}, F^d_{\mathrm{mon}}, \Lambda^{\mathrm{std}}) \ge 2^{d-1}.$$

*Proof.* Let $N$ be any adaptive information mapping. We consider functions $f$ for which we obtain the same information $\mathbf{y} = N(f) = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))$ as for the diagonal split function $\mathbb{1}\left[|\mathbf{x}|_1 \ge \frac{d}{2}\right]$. Consequently, such functions will be evaluated at the same points $\mathbf{x}_1, \dots, \mathbf{x}_n \in [0,1]^d$, and

$$f(\mathbf{x}_i) = \begin{cases} 0 & \text{if } |\mathbf{x}_i|_1 < \frac{d}{2}, \\ 1 & \text{if } |\mathbf{x}_i|_1 \ge \frac{d}{2}, \end{cases}$$

for $i = 1, \dots, n$. Having this information, there are two areas,

$$D_0 := \{\mathbf{x} \in [0,1]^d \mid \exists\, \mathbf{x}_i \ge \mathbf{x}, f(\mathbf{x}_i) = 0\} = \bigcup_{i: f(\mathbf{x}_i)=0} [\![\mathbf{0}, \mathbf{x}_i]\!], \quad \text{and}$$
$$D_1 := \{\mathbf{x} \in [0,1]^d \mid \exists\, \mathbf{x}_i \le \mathbf{x}, f(\mathbf{x}_i) = 1\} = \bigcup_{i: f(\mathbf{x}_i)=1} [\![\mathbf{x}_i, \mathbf{1}]\!],$$

where we know the function for sure: $f|_{D_0} = 0$ and $f|_{D_1} = 1$. The minimal monotone function fitting to that information is $f_-(\mathbf{x}) := \mathbb{1}[\mathbf{x} \notin D_0]$, the maximal function is $f_+(\mathbf{x}) := \mathbb{1}[\mathbf{x} \in D_1]$. Their $L_1$-distance is

$$\|f_+ - f_-\|_1 = \lambda^d\left([0,1]^d \setminus (D_0 \cup D_1)\right),$$

so we have the error bound

$$e(\phi \circ N, F^d_{\mathrm{mon}}) \ge \tfrac{1}{2}\,\lambda^d\left([0,1]^d \setminus (D_0 \cup D_1)\right), \tag{4.2.1}$$

no matter which output $\phi$ we choose. For $|\mathbf{x}_i|_1 < \frac{d}{2}$ we have

$$\lambda^d([\![\mathbf{0}, \mathbf{x}_i]\!]) = \prod_{j=1}^{d} \mathbf{x}_i(j) \overset{\text{AM-GM ineq.}}{\le} \left(\frac{|\mathbf{x}_i|_1}{d}\right)^d < 2^{-d}.$$

For $|\mathbf{x}_i|_1 \ge \frac{d}{2}$ we have a similar estimate,

$$\lambda^d([\![\mathbf{x}_i, \mathbf{1}]\!]) = \prod_{j=1}^{d} (1 - \mathbf{x}_i(j)) \overset{\text{AM-GM ineq.}}{\le} \left(\frac{|\mathbf{1} - \mathbf{x}_i|_1}{d}\right)^d \le 2^{-d}.$$

Consequently,

$$\lambda^d\left([0,1]^d \setminus (D_0 \cup D_1)\right) \geq 1 - n\,2^{-d}\,,$$

which together with (4.2.1) finishes the proof. □

For the integration the proof follows exactly the same lines. While for integration the standard Monte Carlo method easily achieves strong polynomial tractability, for the approximation we still have intractability in the randomized setting, see Section 4.3, yet the curse of dimensionality is broken, see Section 4.4.2.

With slight modifications an analogue lower bound on the deterministic approximation of Boolean monotone functions is found.

**Theorem 4.5.** *For the approximation of monotone Boolean functions we have*

$$e^{\mathrm{det}}(n, \mathrm{APP}, F^d_{\circ|}, \Lambda^{\mathrm{std}}) \geq \frac{1}{2}\,\left(1 - n\,2^{-\lfloor d/2\rfloor}\right)\,.$$

*Hence, for $0 < \varepsilon < \frac{1}{4}$, the $\varepsilon$-complexity is bounded from below by*

$$n^{\mathrm{det}}(\varepsilon, \mathrm{APP}, F^d_{\circ|}, \Lambda^{\mathrm{std}}) \geq 2^{\lfloor d/2\rfloor - 1}\,.$$

*In particular, the approximation of monotone Boolean functions suffers from the curse of dimensionality in the worst case setting.*

*Proof.* Similarly to the proof of Theorem 4.4, we consider fooling functions with

$$f(\mathbf{x}_i) = \begin{cases} 0 & \text{if } |\mathbf{x}_i|_1 < \frac{d}{2}, \\ 1 & \text{if } |\mathbf{x}_i|_1 \geq \frac{d}{2}. \end{cases}$$

For $|\mathbf{x}_i|_1 < \frac{d}{2}$ we have the estimate $\#\{\mathbf{x} \in \{0,1\}^d \mid \mathbf{x} \leq \mathbf{x}_i\} \leq 2^{\lceil d/2\rceil - 1}$, for $|\mathbf{x}_i|_1 \geq \frac{d}{2}$ it holds $\#\{\mathbf{x} \in \{0,1\}^d \mid \mathbf{x} \geq \mathbf{x}_i\} \leq 2^{\lfloor d/2\rfloor}$. The remaining steps are analogous. □

**Remark 4.6** (On the initial error)**.** For the class $F^d_{\mathrm{mon}}$ of real-valued monotone functions, it is easy to see that the initial error is $\frac{1}{2}$ with the initial guess being the constant $\frac{1}{2}$-function.

For Boolean functions, since there is no $\frac{1}{2}$-function, we need to exploit monotonicity properties in order to show that the initial error is $\frac{1}{2}$ nevertheless. Let the initial guess be $g(\mathbf{x}) := x_1$, indeed $g \in F^d_{\circ|}$. For any $f \in F^d_{\circ|}$ we have

$$\begin{aligned} \#\{\mathbf{x} \in \{0,1\}^d \mid f(\mathbf{x}) \neq x_1\} &= \#\{\mathbf{x} \mid f(\mathbf{x}) = 1, x_1 = 0\} + \#\{\mathbf{x} \mid f(\mathbf{x}) = 0, x_1 = 1\} \\ [\text{monotonicity}] \quad &\leq \#\{\mathbf{x} \mid f(\mathbf{x}) = 1, x_1 = 1\} + \#\{\mathbf{x} \mid f(\mathbf{x}) = 0, x_1 = 1\} \\ &= 2^{d-1}\,, \end{aligned}$$

from which we conclude $\mathrm{dist}(f, g) \leq \frac{1}{2}$.

In fact, there are several functions that are equally suitable for an initial guess, a more canonical function without bias to one single coordinate could be the diagonal split $g(\mathbf{x}) := \mathbb{1}[|\mathbf{x}|_1 \geq \frac{d}{2}]$, for Boolean monotone functions, however, this only works properly for odd $d$.

The monotonicity assumption is crucial for us to prove the initial error to be independent from the algorithmic setting. In contrast to that, the approximation of general (non-monotone) Boolean functions with Boolean approximants is an interesting problem where it makes a difference whether we consider the randomized or the deterministic initial error. In the randomized setting, the algorithm that uses no information can return the constant 0 and the constant 1 functions with probability $\frac{1}{2}$ each, that way obtaining the initial error $\frac{1}{2}$. In the deterministic setting, however, for any initial guess $g$, the opposite function $f := \neg g$ has a distance 1, the initial error is 1.

Anyway, with the initial error being constant independently from the dimension $d$, the problem of the approximation of monotone (Boolean) functions is properly normalized. In addition, problems of lower dimensions are canonically contained in the problem of higher dimensions, $F^d_{\mathrm{mon}}$ can be seen as a subset of $F^{d+1}_{\mathrm{mon}}$ consisting of all functions that do not depend on the coordinate $x_{d+1}$. These two properties make the class of multivariate monotone functions particularly interesting for tractability studies.

**Remark 4.7** (A combined lower bound)**.** While the lower bound of Theorem 4.2 contains bad $d$-dependent constants, the bound from Theorem 4.4 does not work for small $\varepsilon$. One could combine the ideas of both proofs in the following way:

Given $\mathbf{m} \in \mathbb{N}^d$, split the domain $[0,1]^d$ into $\prod \mathbf{m} := m_1 \dots m_d$ sub-cuboids $C_{\mathbf{i}}$, each of side length $m_j^{-1}$ in the $j$-th dimension, and indexed by $\mathbf{i} \in \mathbb{N}_0^d$, $i_j \in \{0, \dots, m_j - 1\}$. We consider those monotone functions $f \in F^d_{\mathrm{mon}}$ with different ranges on different sub-cuboids,

$$f(\mathbf{x}) \in \left[ \frac{|\mathbf{i}|_1}{|\mathbf{m}|_1 - d + 1}, \frac{|\mathbf{i}|_1 + 1}{|\mathbf{m}|_1 - d + 1} \right] \quad \text{for } \mathbf{x} \in C_{\mathbf{i}}.$$

Note that, by construction, on each of the sub-cuboids, $f|_{C_{\mathbf{i}}}$ can be chosen independently from the values of the function on other sub-cuboids. If an algorithm computes $n_{\mathbf{i}}$ function values of $f$ on $C_{\mathbf{i}}$, we can apply Theorem 4.4 to the approximation of $f|_{C_{\mathbf{i}}}$ by scaling. Altogether, with $n = \sum_{\mathbf{i}} n_{\mathbf{i}}$, we obtain the lower bound

$$\begin{aligned} e^{\mathrm{det}}(n, \mathrm{APP}, F^d_{\mathrm{mon}}, \Lambda^{\mathrm{std}}) &\geq \frac{1}{|\mathbf{m}|_1 - d + 1} \left[ \frac{1}{\prod \mathbf{m}} \sum_{\mathbf{i}} \tfrac{1}{2} \left(1 - n_{\mathbf{i}}\, 2^{-d}\right) \right] \\ &= \frac{1}{2(|\mathbf{m}|_1 - d + 1)} \left( 1 - \frac{n}{2^d \prod \mathbf{m}} \right). \end{aligned}$$

For $\frac{1}{4(d+1)} \leq \varepsilon \leq \frac{1}{4}$, choose $\mathbf{m} = (2, \dots, 2, 1, \dots, 1)$, splitting the domain only for the first $k$ coordinates, $k := \lfloor \frac{1}{4\varepsilon} \rfloor - 1 \in \{0, \dots, d\}$.
For $0 < \varepsilon \leq \frac{1}{4(d+1)}$, split the domain into $m^d$ subcubes, that is, $\mathbf{m} = (m, \dots, m)$ with $m := \lfloor \frac{1}{4\,\varepsilon\, d} - \frac{1}{d} + 1 \rfloor \geq 2$.
By this we obtain the complexity bound

$$n^{\mathrm{det}}(\varepsilon, \mathrm{APP}, F^d_{\mathrm{mon}}, \Lambda^{\mathrm{std}}) \geq \begin{cases} 2^{d + \lfloor 1/(4\,\varepsilon) \rfloor - 2} & \text{for } \varepsilon \in [\frac{1}{4(d+1)}, \frac{1}{4}], \\ 2^{d\,(1 + \log_2 \lfloor 1/(4\,\varepsilon\,d) \rfloor) - 1} & \text{for } \varepsilon \in (0, \frac{1}{4(d+1)}]. \end{cases}$$

Still, the upper bounds in Theorem 4.2 for the complexity are superexponential in $d$, from Remark 4.3 we have

$$n^{\mathrm{det}}(\varepsilon, \mathrm{APP}, F_{\mathrm{mon}}^d, \Lambda^{\mathrm{std}}) \leq \exp\left(d \log \frac{d}{2\,\varepsilon}\right),$$

there is a logarithmic gap in the exponent.

This idea of a combined lower bound can also be done for the randomized setting, see Remark 4.16.

## 4.3 Intractability for Randomized Approximation

### 4.3.1 The Result – A Monte Carlo Lower Bound

As we will show in Section 4.4, for the $L_1$-approximation of monotone functions, the curse of dimensionality does not hold anymore in the randomized setting. Within this section, however, we show that, for fixed $\varepsilon \in (0, \frac{1}{2})$, the $\varepsilon$-complexity depends at least exponentially on $\sqrt{d}$ in the randomized setting. Yet worse, the problem is *not weakly tractable*.

For the proof we switch to an average case setting for Boolean functions, an idea that has already been used by Blum, Burch, and Langford [8, Sec 4].[5] They stated that for $n \geq d$, and *sufficiently large* $d$, we have

$$e^{\mathrm{ran}}(n, \mathrm{APP}, F_{\circ|}^d, \Lambda^{\mathrm{std}}) \geq \left(1 - \exp\left(-\tfrac{n}{4}\right)\right) \left(\frac{1}{2} - C\,\frac{\log(6\,d\,n)}{\sqrt{d}}\right), \tag{4.3.1}$$

with a numerical constant $C > 0$. (From their proof the constant $C = 2.9895...$ can be extracted.) Assuming this result to hold for all $n$ and $d$, one could conclude that for $n = \lfloor \frac{1}{6\,d} \exp(\sigma\sqrt{d}) \rfloor$ we have a lower error bound of roughly[6] $\varepsilon = \frac{1}{2} - C\sigma$, where we may choose $0 < \sigma < \frac{1}{2\,C}$ for meaningful bounds. Consequently, such an estimate gives us results only for $n < \frac{1}{6\,d} \exp(\sqrt{d}/(2\,C))$. With the given value of $C$, this means that $d = 8799$ is the smallest dimension for which a positive error bound with $n = 1$ is possible in the first place.[7] Actually, interpreting "sufficiently large" as $d \geq 8799$, one can check their proof and will find out that all proof steps work in those cases where we have non-trivial error bounds, see Remark 4.15 for more information on the original proof.

Note that

$$\frac{1}{d} \exp(\sigma\,\sqrt{d}) \succ \exp(\sigma'\,\sqrt{d}) \quad \text{for } d \to \infty, \text{ if } 0 < \sigma' < \sigma.$$

Therefore, by (4.3.1), Blum et al. were the first to show that for fixed $\varepsilon \in (0, \frac{1}{2})$ the Monte Carlo complexity for the approximation of monotone Boolean functions depends at least exponentially on $\sqrt{d}$.

[5] I wish to thank Dr. Mario Ullrich for pointing me to this paper, after learning about my first version of a lower bound proof.

[6] That is, ignoring the prefactor $(1 - \exp(-\frac{n}{4}))$.

[7] One more example: $d = 39\,168$ is the first dimension for which a positive lower bound with $n \geq d$ is possible.

From the IBC point of view the structure of (4.3.1) appears unfamiliar. Blum et al., however, wanted to show that if we allow to use algorithms with cardinality $n$ growing only polynomially in the dimension $d$, the error approaches the initial error with a rate of almost $1/\sqrt{d}$. The interest for such a result is *not* motivated from practically *solving* the problem, but from proving that the problem is *hard to solve*, and could therefore be used in cryptographic procedures.[8]

One objective in this study was to extract the best constants possible whenever a constant error tolerance $\varepsilon$ is given. Moreover, using some different inequalities than Blum et al.,[9] it is possible to find a lower complexity bound that includes the error tolerance $\varepsilon$ in a way that we can prove *intractability* for the Monte Carlo approximation of monotone functions, see Remark 4.9. A detailed comment on modifications of the proof needed for the case of $\varepsilon$ getting close to the initial error, that is the case Blum et al. [8] studied, is made in Remark 4.15.

**Theorem 4.8.** *Consider the randomized approximation of monotone Boolean functions. There exist constants $\sigma_0, \nu, \varepsilon_0 > 0$ and $d_0 \in \mathbb{N}$ such that for $d \geq d_0$ we have*

$$n^{\mathrm{ran}}(\varepsilon_0, \mathrm{APP}, F_{\circ|}^d, \Lambda^{\mathrm{std}}) > \nu \, \exp(\sigma_0 \sqrt{d}) \,,$$

*and moreover, for $0 < \varepsilon \leq \varepsilon_0$ and $d \geq d_0 \, (\varepsilon_0/\varepsilon)^2$ we have*

$$n^{\mathrm{ran}}(\varepsilon, \mathrm{APP}, F_{\circ|}^d, \Lambda^{\mathrm{std}}) > \nu \, \exp\left(\frac{c\sqrt{d}}{\varepsilon}\right) ,$$

*with $c = \sigma_0 \, \varepsilon_0$ .*

*In particular, for $d \geq d_0 = 100$ and $\varepsilon_0 = \frac{1}{30}$ we have*

$$n^{\mathrm{ran}}(\tfrac{1}{30}, \mathrm{APP}, F_{\circ|}^d, \Lambda^{\mathrm{std}}) > 108 \cdot \exp(\sqrt{d} - \sqrt{100}) \,,$$

*for $d = 100$ this means $n^{\mathrm{ran}} > 108$. For $0 < \varepsilon \leq \frac{1}{30}$ and $d \geq 1/(9\,\varepsilon^2)$ we have*

$$n^{\mathrm{ran}}(\varepsilon, \mathrm{APP}, F_{\circ|}^d, \Lambda^{\mathrm{std}}) > 108 \cdot \exp\left(\frac{\sqrt{d}}{30\,\varepsilon} - \sqrt{100}\right) .$$

*All these lower bounds hold for varying cardinality as well.*

[8] Think of two parties **A** and **B** communicating. Both have a key, say, they know a high-dimensional Boolean function $f$. In order to check the identity of the other party they ask for some values of $f$. That is, **A** sends a list of points $\mathbf{x}_1, \ldots, \mathbf{x}_k$ to **B**, and **B** answers with a list of correct values $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_k)$ in order to approve a message. A third party **C** with bad intentions could intercept this communication and try to learn the function $f$ from sample pairs $(\mathbf{x}_i, f(\mathbf{x}_i))$. This should be as hard as possible for that there is little chance that one day **C** could answer correctly to $\mathbf{x}_1, \ldots, \mathbf{x}_k$.
Monotone Boolean functions are just one model of such a problem. Besides *hardness of learning*, another criterion is *simplicity of representation* (keeping the key), and probably other problems are better for that purpose.
The big issue for complexity theory: With technical progress hackers have more capacities for cracking a code, with more intense communication they can collect more information. But technical progress also gives opportunities to make cryptography more complex in order to rule out cyber attacks.

[9] Differences within the proof will be indicated in place by footnotes.

Before we give the proof in Section 4.3.2, we discuss some consequences of the theorem.

**Remark 4.9** (Intractability)**.** The above theorem shows that the approximation of monotone Boolean functions is *not weakly tractable*. Indeed, consider the sequence $(\varepsilon_d)_{d=d_0}^{\infty}$ of error tolerances $\varepsilon_d := \varepsilon_0 \sqrt{d_0/d}$. Then

$$\lim_{\varepsilon_d^{-1}+d\to\infty} \frac{\log n(\varepsilon_d, d)}{\varepsilon_d^{-1} + d} \geq \lim_{d\to\infty} \frac{\sigma_0\, d/\sqrt{d_0} + \log \nu}{\varepsilon_0^{-1}\sqrt{d/d_0} + d} = \frac{\sigma_0}{\sqrt{d_0}} > 0\,.$$

This contradicts the definition of weak tractability, see Section 1.2.

Actually, this behaviour has already been known since the paper of Bshouty and Tamon 1996 [9, Thm 5.3.1], however, research on weak tractability has not yet been started at that time.[10] Their lower bound can be summarized as follows: For moderately decaying error tolerances $\varepsilon_d \preceq 1/(\sqrt{d}(1+\log d))$ and sufficiently large $d$, we have

$$n^{\mathrm{ran}}(\varepsilon_d, \mathrm{APP}, F_{\circ|}^d, \Lambda^{\mathrm{std}}) \geq c\, 2^d/\sqrt{d}\,,$$

with some numerical constant $c > 0$. Interestingly, the proof is based on purely combinatorial arguments, without applying Bakhvalov's technique (Proposition 1.2) and average case settings. Since a function value for Boolean functions may only be 0 or 1, after $n$ oracle calls we have at most $2^n$ different possible information outcomes. Any Boolean function can approximate at most

$$k(\varepsilon, d) := \sum_{l=0}^{\lfloor \varepsilon\, 2^d \rfloor} \binom{2^d}{l} \overset{\text{Lem B.1}}{\leq} 2^{2^d\, \varepsilon\, \log_2(\mathrm{e}/\varepsilon)}$$

Boolean functions up to a distance $\varepsilon \in (0, 1]$. On the other hand, the total number of monotone Boolean functions is known to be

$$\#F_{\circ|}^d \geq 2^{\binom{d}{\lfloor d/2 \rfloor}} \geq 2^{2^{d-1}/\sqrt{d}}\,,$$

see Korshunov [37, Sec 1.1] and the references therein for the first inequality, and Lemma B.2 for the second inequality. Then for any realization of a Monte Carlo method, that is, fixing $\omega$, a portion of at least $(1 - 2^n\, k(\varepsilon, d)/\#F_{\circ|}^d)$ Boolean monotone functions is at distance more than $\varepsilon$ to all of the output functions. This can be used to show the existence of poorly approximated functions,

$$\begin{aligned}
\sup_{f\in F_{\circ|}^d} \mathbb{P}\{\mathrm{dist}(A_n^\omega, f) > \varepsilon\} &\geq \frac{1}{\#F_{\circ|}^d} \sum_{f\in F_{\circ|}^d} \mathbb{P}\{\mathrm{dist}(A_n^\omega, f) > \varepsilon\} \\
[\text{Fubini}] \quad &= \mathbb{E}\, \frac{\#\{f \in F_{\circ|}^d \mid \mathrm{dist}(A_n^\omega, f) > \varepsilon\}}{\#F_{\circ|}^d} \\
&\geq \left(1 - 2^n\, \frac{k(\varepsilon, d)}{\#F_{\circ|}^d}\right).
\end{aligned}$$

[10] For the historical background of tractability notions, see Novak and Woźniakowski 2008 [55, p. 9].

Hence we get the error bound

$$e^{\mathrm{ran}}(n, \mathrm{APP}, F^d_{\circ|}, \Lambda^{\mathrm{std}}) \geq \varepsilon \left(1 - 2^n \, \frac{k(\varepsilon, d)}{\#F^d_{\circ|}}\right),$$

which implies the complexity bound

$$\begin{aligned} n^{\mathrm{ran}}(\tfrac{\varepsilon}{2}, \mathrm{APP}, F^d_{\circ|}, \Lambda^{\mathrm{std}}) &\geq \log_2 \left(\frac{\#F^d_{\circ|}}{2\,k(\varepsilon, d)}\right) \\ &\geq \frac{2^{d-1}}{\sqrt{d}} - 2^d\,\varepsilon\,\log_2\left(\frac{\mathrm{e}}{\varepsilon}\right) - 1\,. \end{aligned}$$

This lower bound only makes sense for $\varepsilon \log \varepsilon^{-1} \preceq 1/\sqrt{d}$ as $d \to \infty$, then it will give a complexity bound that is exponential in $d$.

**Remark 4.10** (Real-valued monotone functions)**.** The lower bounds of Theorem 4.8 also hold for the problem $\mathrm{APP} : F^d_{\mathrm{mon}} \hookrightarrow L_1[0,1]^d$, which therefore is intractable as well.

Indeed, the proof of the theorem is done by switching to a $\mu$-average case setting on the set of monotone Boolean functions $F^d_{\circ|}$ (Bakhvalov's technique, see Proposition 1.2). Any measure $\mu$ on $F^d_{\circ|}$ can be associated with a measure $\tilde{\mu}$ on the set of subcubewise constant functions $\widetilde{F}^d_{\circ|} \subset F^d_{\mathrm{mon}}$ that only take the values 0 and 1. Since for the Boolean setting the output function must be Boolean as well, in order to prove the equivalence of both problems, it remains to show that optimal outputs with respect to $\tilde{\mu}$ are constant 0 or 1 on each of the $2^d$ subcubes of the domain $[0,1]^d$.

For the $\tilde{\mu}$-average setting on $F^d_{\mathrm{mon}}$, take any deterministic information mapping $\widetilde{N} : f \mapsto \mathbf{y} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) \in \{0,1\}^n$ using $n$ sample points (possibly adaptively chosen). By $\tilde{\mu}_{\mathbf{y}}(\cdot) = \tilde{\mu}(\cdot \mid \widetilde{N}(f) = \mathbf{y})$ we denote the conditional measure, and $F_{\mathbf{y}} := \widetilde{N}^{-1}(\mathbf{y}) \subseteq F^d_{\mathrm{mon}}$ is the preimage of the information $\mathbf{y}$. We only need to consider the cases of information $\mathbf{y}$ with non-vanishing probability $\tilde{\mu}(F_{\mathbf{y}})$. Taking any output mapping $\tilde{\phi}$, the definition of the average error and the law of total probability lead to the following representation of the error:

$$\begin{aligned} e(\tilde{\phi} \circ \widetilde{N}, \tilde{\mu}) \;&=\; \sum_{\mathbf{y}} \tilde{\mu}(F_{\mathbf{y}}) \int_{F_{\mathbf{y}}} \big\| [\tilde{\phi}(\mathbf{y})](\mathbf{x}) - f(\mathbf{x}) \big\|_{L_1} \tilde{\mu}_{\mathbf{y}}(\mathrm{d}f) \\ &\overset{\text{Fubini}}{=} \sum_{\mathbf{y}} \tilde{\mu}(F_{\mathbf{y}}) \int_{[0,1]^d} \left[ \int_{F_{\mathbf{y}}} \big| [\tilde{\phi}(\mathbf{y})](\mathbf{x}) - f(\mathbf{x}) \big| \tilde{\mu}_{\mathbf{y}}(\mathrm{d}f) \right] \mathrm{d}\mathbf{x} \end{aligned} \tag{4.3.2}$$

Since $f(\mathbf{x}) \in \{0,1\}$, for the integrand we have

$$[\ldots] \;=\; \tilde{\mu}_{\mathbf{y}}\{f(\mathbf{x}) = 0\} \, \big| [\tilde{\phi}(\mathbf{y})](\mathbf{x}) \big| + \tilde{\mu}_{\mathbf{y}}\{f(\mathbf{x}) = 1\} \, \big| 1 - [\tilde{\phi}(\mathbf{y})](\mathbf{x}) \big|\,,$$

which is minimized for $[\tilde{\phi}(\mathbf{y})](\mathbf{x}) := \mathbb{1}\big[\tilde{\mu}_{\mathbf{y}}\{f(\mathbf{x}) = 1\} \geq \frac{1}{2}\big]$. This, of course, is a function that is constant 0 or 1 on each of the $2^d$ subcubes.

Note that, since the set of Boolean functions is finite, by minimax principles there exists a measure $\mu^*$ on $F^d_{\circ|}$ such that the $\mu^*$-average error coincides with the Monte

Carlo error, see Mathé [47]. By this we have that the problem of approximating Boolean functions is strictly easier than the problem of $L_1$-approximation of real-valued functions,

$$e^{\mathrm{ran}}(n, \mathrm{APP}, F^d_{\circ|}, \Lambda^{\mathrm{std}}) = e^{\mathrm{avg}}(n, \mathrm{APP}, \mu^*, \Lambda^{\mathrm{std}}) \le e^{\mathrm{ran}}(n, \mathrm{APP}, F^d_{\mathrm{mon}}, \Lambda^{\mathrm{std}})\,.$$

### 4.3.2 The Proof of the Monte Carlo Lower Bound

We start the proof with two preparatory lemmas.

The calculation (4.3.2) in Remark 4.10 actually brings us to a direct representation of the best error possible with the information $\widetilde{N}$,

$$\inf_{\tilde\phi} e(\tilde\phi \circ \widetilde{N}, \tilde\mu) = \sum_{\mathbf{y}} \tilde\mu(F_{\mathbf{y}}) \int_{[0,1]^d} \min\bigl\{\tilde\mu_{\mathbf{y}}\{f(\mathbf{x}) = 0\}, \tilde\mu_{\mathbf{y}}\{f(\mathbf{x}) = 1\}\bigr\}\mathrm{d}\mathbf{x}\,.$$

We summarize a similar identity for Boolean functions in the following Lemma.

**Lemma 4.11** (Error for the optimal output)**.** *Let $\mu$ be a probability measure on the set $F^d_{\circ|}$, and let $N : F^d_{\circ|} \to \{0,1\}^n$ be any deterministic information mapping. Then*

$$\inf_{\phi} e(\phi \circ N, \mu) = \sum_{\mathbf{y}} \mu(F_{\mathbf{y}})\, 2^{-d} \sum_{\mathbf{x}\in\{0,1\}^d} \min\bigl\{\mu_{\mathbf{y}}\{f(\mathbf{x}) = 0\}, \mu_{\mathbf{y}}\{f(\mathbf{x}) = 1\}\bigr\}\,, \quad (4.3.3)$$

*where $\mu_{\mathbf{y}}(\cdot) = \mu(\,\cdot \mid N(f) = \mathbf{y})$ is the conditional measure, and $F_{\mathbf{y}} := N^{-1}(\mathbf{y}) \subseteq F^d_{\circ|}$ is the preimage of the information $\mathbf{y}$.*

By Lemma 4.11, for a given measure $\mu$ on $F^d_{\circ|}$, the average case analysis reduces to understanding the conditional measure $\mu_{\mathbf{y}}$. The concept of *augmented information* allows us to simplify the conditional measure as long as we are concerned about lower bounds.

**Lemma 4.12** (Augmented information)**.** *Consider the general problem $S : F \to G$. Let $N : F \to \mathcal{Y}$ be an arbitrary measurable information mapping and $\widetilde{N} : F \to \widetilde{\mathcal{Y}}$ be an* augmented information mapping*, that is, for all possible information representers $\tilde y \in \widetilde{\mathcal{Y}}$ there exists an information representer $\mathbf{y} \in \mathcal{Y}$ such that $\widetilde{N}^{-1}(\tilde y) \subseteq N^{-1}(\mathbf{y})$.*

*With this property, the augmented information allows for smaller errors, that means*

$$\inf_{\tilde\phi} e(\tilde\phi \circ \widetilde{N}, \mu) \le \inf_{\phi} e(\phi \circ N, \mu)\,.$$

*Proof.* There exists a mapping $\psi : \widetilde{\mathcal{Y}} \to \mathcal{Y}$ such that $N = \psi \circ \widetilde{N}$. For any mapping $\phi : \mathcal{Y} \to G$ we can define $\tilde\phi := \phi \circ \psi$ such that $\phi \circ N = \phi \circ \psi \circ \widetilde{N} = \tilde\phi \circ \widetilde{N}$, and so

$$r^{\mathrm{avg}}(\widetilde{N}, \mu) := \inf_{\tilde\phi} e(\tilde\phi \circ \widetilde{N}, \mu) \le \inf_{\phi} e(\phi \circ N, \mu) =: r^{\mathrm{avg}}(N, \mu)\,.$$

(The quantity $r^{\mathrm{avg}}(N, \mu)$ is called *$\mu$-average radius* of the information $N$.) □

We are now ready to proof the theorem.

*Proof of Theorem 4.8.* We will use Bakhvalov's technique (Proposition 1.2) for lower bounds and switch to an average case setting on the set of monotone Boolean functions. The proof is organized in seven steps.

**Step 1:** The general structure of the measure $\mu$ on $F_{\circ\mid}^d$.

**Step 2:** Introduce the augmented information.

**Step 3:** Estimate the number of points $\mathbf{x} \in \{0,1\}^d$ for which $f(\mathbf{x})$ is still – to some extend – undetermined, even after knowing the augmented information.

**Step 4:** Further specify the measure $\mu$, and give estimates on the conditional probability for the event $f(\mathbf{x}) = 0$ for the set of still fairly uncertain $\mathbf{x}$ from the step before.

**Step 5:** A general formula for the lower bound.

**Step 6:** Connect estimates for $\varepsilon_0$ and $d_0$ with estimates for smaller $\varepsilon$ and larger $d$.

**Step 7:** Explicit numerical values.

**Step 1: General structure of the measure $\mu$.**
We define a measure $\mu$ on the set of functions that can be represented by a randomly drawn set $U \subseteq W := \{\mathbf{x} \in \{0,1\}^d \mid |\mathbf{x}|_1 = t\}$, with $t \in \mathbb{N}$ being a suitable parameter, and a boundary value $b \in \mathbb{N}$, $t \le b \le d$. We define $f_U : \{0,1\}^d \to \{0,1\}$ by

$$f_U(\mathbf{x}) := \mathbb{1}\left[(|\mathbf{x}|_1 > b) \text{ or } (\exists \mathbf{u} \in U : \mathbf{u} \le \mathbf{x})\right] . \tag{4.3.4}$$

The boundary value $b \in \mathbb{N}$ will facilitate considerations in connection with the augmented information.[11] We draw $U$ such that the $f(\mathbf{w})$ are independent Bernoulli random variables with $p = \mu\{f(\mathbf{w}) = 1\} = 1 - \mu\{f(\mathbf{w}) = 0\}$. The parameter $p \in (0,1)$ will be specified in Step 4.

[11] The proof of Blum et al. [8] worked without a boundary value $b$ when defining functions $f_U$ for the measure $\mu$. Then in Step 2 one needs to replace the first inequality of (4.3.6). They used the Chernoff bound in order to control the size of the set $V_0$ of the augmented information $\tilde{y}$ with high probability,

$$\mu\{\#V_0 \le \tfrac{2n}{p}\} \ge 1 - \exp(-\tfrac{n}{4}) .$$

The error bound, see Step 5, then needs to be multiplied with this factor, which for large $n$ is neglectable, though. In Step 4 we specify $p = \varrho/\binom{a}{t}$, so we could use the estimate $\#V_0 \le n\,\frac{2}{\varrho}\binom{a}{t}$. The estimate (4.3.6) we take instead can be compared to this using (4.3.13), we have $\#V_0 \le n\binom{b}{t} \le n\,\sigma_{\alpha\beta\tau}(d)\binom{a}{t}$. For high dimensions we get the rough estimate $\sigma_{\alpha\beta\tau}(d) \approx \exp((\beta-\alpha)\tau)$. With the numerical values listed in Step 7 we have $\frac{2}{\varrho} = 7.7041...$ versus $\sigma_{\alpha\beta\tau}(d) = 6.5622...$. In this case our version is slightly better and avoids the usage of Chernoff bounds.

For the original result, however, Chernoff bounds prove to be useful, see Remark 4.15. In turn, the case of varying cardinality becomes more difficult because the situation of Lemma 1.6 does not apply anymore, compare Step 5.

**Step 2: Augmented information.**
Now, for any (possibly adaptively obtained) info $\mathbf{y} = N(f) = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))$ with $\mathbf{x}_i \in \{0,1\}^d$, we define the augmented information

$$\tilde{y} := (V_0, V_1), \tag{4.3.5}$$

where $V_0 \subseteq W \setminus U$ and $V_1 \subseteq U$ represent knowledge about the instance $f$ that implies the information $\mathbf{y}$. We know $f(\mathbf{u}) = 0$ for $\mathbf{u} \in V_0$, and $f(\mathbf{u}) = 1$ for $\mathbf{u} \in V_1$. In detail, let $\leq_{\mathrm{L}}$ be the lexicographic order[12] of the elements of $W$, then $\min_{\mathrm{L}} V$ denotes the first element of a set $V \subseteq W$ with respect to this order. For a single sample $f(\mathbf{x})$ the augmented oracle reveals the sets

$$V_0^{\mathbf{x}} := \begin{cases} \emptyset & \text{if } |\mathbf{x}|_1 > b, \\ \{\mathbf{v} \in W \mid \mathbf{v} \leq \mathbf{x}\} & \text{if } f(\mathbf{x}) = 0, \\ \{\mathbf{v} \in W \mid \mathbf{v} \leq \mathbf{x} \text{ and } \mathbf{v} <_{\mathrm{L}} \min_{\mathrm{L}}\{\mathbf{u} \in U \mid \mathbf{u} \leq \mathbf{x}\}\} & \text{if } f(\mathbf{x}) = 1 \\ & \text{and } |\mathbf{x}|_1 \leq b, \end{cases}$$

$$V_1^{\mathbf{x}} := \begin{cases} \emptyset & \text{if } |\mathbf{x}|_1 > b \text{ or } f(\mathbf{x}) = 0, \\ \{\min_{\mathrm{L}}\{\mathbf{u} \in U \mid \mathbf{u} \leq \mathbf{x}\}\} & \text{if } f(\mathbf{x}) = 1 \text{ and } |\mathbf{x}|_1 \leq b, \end{cases}$$

and altogether the augmented information is

$$\tilde{y} = (V_0, V_1) := \left( \bigcup_{i=1}^{n} V_0^{\mathbf{x}_i}, \bigcup_{i=1}^{n} V_1^{\mathbf{x}_i} \right).$$

Note that computing $f(\mathbf{x})$ for $|\mathbf{x}|_1 > b$ is a waste of information, so no algorithm designer would decide to compute such samples. Since $\#V_0^{\mathbf{x}} \leq \binom{|\mathbf{x}|_1}{t} \leq \binom{b}{t}$ for $|\mathbf{x}|_1 \leq b$, and $\#V_1^{\mathbf{x}} \leq 1$, we have the estimates[11]

$$\#V_0 \leq n \binom{b}{t}, \quad \text{and} \quad \#V_1 \leq n. \tag{4.3.6}$$

**Step 3: Number of points $\mathbf{x} \in \{0,1\}^d$ where $f(\mathbf{x})$ is still fairly uncertain.**
For any point $\mathbf{x} \in \{0,1\}^d$ we define the set

$$W_{\mathbf{x}} := \{\mathbf{w} \in W \mid \mathbf{w} \leq \mathbf{x}\}$$

of points that are "relevant" to $f(\mathbf{x})$. Given an augmented information $\tilde{y} = (V_0, V_1)$, we are interested in points where it is not yet clear whether $f(\mathbf{x}) = 1$ or $f(\mathbf{x}) = 0$. In detail, these are points $\mathbf{x}$ where $W_{\mathbf{x}} \cap V_1 = \emptyset$, for that $f(\mathbf{x}) = 0$ be still possible. Furthermore, $W_{\mathbf{x}} \setminus V_0$ shall be big enough, say $\#(W_{\mathbf{x}} \setminus V_0) \geq M$ with $M \in \mathbb{N}$, so that the conditional probability $p_{\mathbf{x}} := \mu_{\tilde{y}}\{f(\mathbf{x}) = 1\}$ is not too small. For our estimates it will be necessary to restrict to points $|\mathbf{x}|_1 \geq a \in \mathbb{N}$, we suppose $t \leq a \leq b$. The set of all these points will be denoted by

$$B := \{\mathbf{x} \in D_{ab} \mid W_{\mathbf{x}} \cap V_1 = \emptyset,\ \#(W_{\mathbf{x}} \setminus V_0) \geq M\},$$
$$\text{where} \quad D_{ab} := \{\mathbf{x} \in \{0,1\}^d \mid a \leq |\mathbf{x}|_1 \leq b\}.$$

[12] Any other total order will be applicable as well.

We aim to find a lower bound for the cardinality of $B$.

*Step 3.1: Bounding $\#D_{a,b}$.*
Let $a := \lceil \frac{d}{2} + \alpha\frac{\sqrt{d}}{2} \rceil$ and $b := \lfloor \frac{d}{2} + \beta\frac{\sqrt{d}}{2} \rfloor$ with $\alpha < \beta$, then by the Berry-Esseen inequality (on the speed of convergence of the Central Limit Theorem), see Proposition B.3 and in particular Corollary B.4, we have

$$\frac{\#D_{ab}}{\#\{0,1\}^d} = \frac{1}{2^d}\sum_{k=a}^{b}\binom{d}{k} \geq \underbrace{\Phi(\beta)-\Phi(\alpha)}_{=:C_{\alpha\beta}} - \frac{2\,C_0}{\sqrt{d}} =: r_0(\alpha,\beta,d)\,. \tag{4.3.7}$$

Here, $\Phi$ is the cumulative distribution function of standard Gaussian variables.[13]

*Step 3.2: The influence of $\mathbf{w} \in W$ (in particular $\mathbf{w} \in V_1$).*[14]
Now, let $t := \lceil \tau\sqrt{d} \rceil$ with $\tau > 0$, and for $\mathbf{w} \in W$ define

$$Q_{\mathbf{w}} := \{\mathbf{x} \in \{0,1\}^d \mid \mathbf{w} \leq \mathbf{x}\}\,,$$

this is the set of all points inside the area of influence of $\mathbf{w}$. Applying Corollary B.4 once more, we obtain

$$\begin{aligned}
\frac{\#(Q_{\mathbf{w}} \cap D_{ab})}{\#Q_{\mathbf{w}}} &= \frac{\#\{\mathbf{x} \in \{0,1\}^{d-t} \mid a-t \leq |\mathbf{x}|_1 \leq b-t\}}{2^{d-t}} \\
&= \frac{1}{2^{d-t}}\sum_{k=a-t}^{b-t}\binom{d-t}{k} \\
\text{[Cor B.4]} \quad &\leq \Phi\left(\frac{2b-t}{\sqrt{d-t}}\right) - \Phi\left(\frac{2a-t}{\sqrt{d-t}}\right) + \frac{2\,C_0}{\sqrt{d-t}} \\
\text{[(4.3.9), (4.3.10)]} \quad &\leq \left[\underbrace{\Phi\left(\beta-\tau\right)-\Phi\left(\alpha-\tau\right)}_{=:C_{\alpha\beta\tau}} + \underbrace{\left(\frac{1}{\sqrt{2\pi}}+2\,C_0\right)}_{=:C_1}\frac{1}{\sqrt{d}}\right]\frac{1}{\sqrt{1-t/d}}\,,
\end{aligned} \tag{4.3.8}$$

where $1 \leq 1/\sqrt{1-t/d} \leq 1/\sqrt{1-\tau/\sqrt{d}-1/d} =: \kappa_\tau(d)$ for $\tau < \sqrt{d} - 1/\sqrt{d}$, and this factor converges, $\kappa_\tau(d) \xrightarrow[d\to\infty]{} 1$. Within the above calculation, we exploited that the density of the Gaussian distribution is decreasing with growing distance to 0, in

[13] The original proof of Blum et al. [8] uses Hoeffding bounds

$$\frac{\#D_{ab}}{\#\{0,1\}^d} \geq 1 - \exp(-\alpha^2) - \exp(-\beta^2)\,.$$

The Berry-Esseen inequality enables us to obtain better constants. Moreover, for the considerations on small $\varepsilon$ in Step 6, we need to take $\alpha, \beta \to 0$, but this cannot be done with the Hoeffding bound, where for $|\alpha|, |\beta| < \sqrt{\log 2}$ we would have trivial negative bounds.

[14] This step helps to get better constants, but it becomes essential for small $\varepsilon$ in Step 6. For the focus of Blum et al. [8] with $\varepsilon$ being close to the initial error, the estimate $\#(Q_{\mathbf{w}} \cap D_{ab})/\#Q_{\mathbf{w}} \leq 1$ will be sufficient.

detail, for $t_0 < t_1$ and $\kappa \geq 1$ we have

$$\underline{\underline{\Phi(\kappa\, t_1) - \Phi(\kappa\, t_0)}} \;=\; \frac{1}{\sqrt{2\pi}} \int_{\kappa\, t_0}^{\kappa\, t_1} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t \;=\; \frac{\kappa}{\sqrt{2\pi}} \int_{t_0}^{t_1} \exp\left(-\frac{\kappa^2\, s^2}{2}\right) \mathrm{d}s$$
$$\leq\; \frac{\kappa}{\sqrt{2\pi}} \int_{t_0}^{t_1} \exp\left(-\frac{s^2}{2}\right) \mathrm{d}s \;=\; \underline{\underline{\kappa\, [\Phi(t_1) - \Phi(t_0)]}}\,. \quad (4.3.9)$$

Namely, we took $\kappa = 1/\sqrt{1 - t/d}$ which comes from replacing $1/\sqrt{d - t}$ by $1/\sqrt{d}$. Furthermore, we shifted the $\Phi$-function, knowing its derivative being bounded between 0 and $1/\sqrt{2\pi}$, so for $t_0 < t_1$ and $\delta \in \mathbb{R}$ we have

$$\Big|\big[\Phi(t_1 + \delta) - \Phi(t_0 + \delta)\big] - \big[\Phi(t_1) - \Phi(t_0)\big]\Big| \leq \frac{|\delta|}{\sqrt{2\pi}}\,, \qquad (4.3.10)$$

in our case $\delta = t/\sqrt{d} - \tau \leq 1/\sqrt{d}$.

*Step 3.3: The influence of $V_0$.*
Markov's inequality gives us

$$\sum_{\mathbf{w} \in V_0} \#(Q_{\mathbf{w}} \cap D_{ab}) = \sum_{\mathbf{x} \in D_{ab}} \#(W_{\mathbf{x}} \cap V_0) \;\geq\; N\, \#\{\mathbf{x} \in D_{ab} \mid \#(W_{\mathbf{x}} \cap V_0) \geq N\}\,, \qquad (4.3.11)$$

with $N \in \mathbb{N}$. Using this, we can carry out the estimate

$$\begin{aligned}
\#\{\mathbf{x} \in D_{ab} \mid \#(W_{\mathbf{x}} \setminus V_0) \geq M\} \;&=\; \#\{\mathbf{x} \in D_{ab} \mid \#(W_{\mathbf{x}} \cap V_0) \leq \#W_{\mathbf{x}} - M\} \\
&\geq\; \#\{\mathbf{x} \in D_{ab} \mid \#(W_{\mathbf{x}} \cap V_0) \leq \tbinom{a}{t} - M\} \\
&=\; \#D_{ab} - \#\{\mathbf{x} \in D_{ab} \mid \#(W_{\mathbf{x}} \cap V_0) > \tbinom{a}{t} - M\} \\
[(4.3.11)] \quad &\geq\; \#D_{ab} - \frac{1}{\binom{a}{t} - M + 1} \sum_{\mathbf{w} \in V_0} \#(Q_{\mathbf{w}} \cap D_{ab})\,.
\end{aligned} \qquad (4.3.12)$$

*Step 3.4: Final estimates on $\#B$.*
Putting all this together, we estimate the cardinality of $B$:

$$\begin{aligned}
\frac{\#B}{\#\{0,1\}^d} &= \frac{\#\left(\{\mathbf{x} \in D_{ab} \mid \#(W_{\mathbf{x}} \setminus V_0) \geq M\} \setminus \bigcup_{\mathbf{w} \in V_1} Q_{\mathbf{w}}\right)}{\#\{0,1\}^d} \\
[(4.3.12), \text{ any } \mathbf{w} \in W] \quad &\geq \frac{\#D_{ab}}{\#\{0,1\}^d} \\
&\quad - \frac{\#Q_{\mathbf{w}}}{\#\{0,1\}^d} \left(\frac{\#V_0}{\binom{a}{t} - M + 1} + \#V_1\right) \frac{\#(Q_{\mathbf{w}} \cap D_{ab})}{\#Q_{\mathbf{w}}} \\
[(4.3.6), (4.3.7), (4.3.8)] \quad &\geq C_{\alpha\beta} - \frac{2\, C_0}{\sqrt{d}} \\
&\quad - n\, 2^{-t} \left(\frac{\binom{b}{t}}{\binom{a}{t} - M + 1} + 1\right) \left[C_{\alpha\beta\tau} + \frac{C_1}{\sqrt{d}}\right] \kappa_\tau(d)\,.
\end{aligned}$$

We set $M := \lceil \lambda \binom{a}{t} \rceil$ with $0 < \lambda < 1$, and provided $t \le a \le b$, which can be guaranteed for $\alpha - 2\tau > -\sqrt{d} + 2/\sqrt{d}$ and $\beta - \alpha > 2/\sqrt{d}$, we estimate the ratio

$$\begin{aligned}\binom{b}{t} \Big/ \binom{a}{t} &\le \left(\frac{a+1}{a-t+1}\right)^{b-a} \le \left(\frac{\frac{d}{2} + \alpha\frac{\sqrt{d}}{2} + 1}{\frac{d}{2} + (\alpha - 2\tau)\frac{\sqrt{d}}{2}}\right)^{(\beta-\alpha)\sqrt{d}/2} \\ &\le \exp\Bigg((\beta-\alpha)\,\tau \underbrace{\left(1 + \frac{\alpha - 2\tau}{\sqrt{d}}\right)^{-1}}_{=:\kappa_{\alpha\tau}(d)} + \underbrace{\frac{\beta-\alpha}{\sqrt{d} + \alpha - 2\tau}}_{=:K_{\alpha\beta\tau}(d)}\Bigg) =: \sigma_{\alpha\beta\tau}(d)\,,\end{aligned} \tag{4.3.13}$$

where we have $1 \le \kappa_{\alpha\tau}(d) \xrightarrow[d\to\infty]{} 1$ and $0 \le K_{\alpha\beta\tau}(d) \xrightarrow[d\to\infty]{} 0$. (Note that the above estimate is asymptotically optimal, $1 \le \binom{b}{t}/\binom{a}{t} \xrightarrow[d\to\infty]{} \exp\left((\beta-\alpha)\,\tau\right)$.) We finally choose the information cardinality $n = \lfloor \nu 2^t \rfloor$, and obtain the estimate

$$\begin{aligned}\frac{\#B}{\#\{0,1\}^d} &\ge \left[C_{\alpha\beta} - \frac{2\,C_0}{\sqrt{d}}\right] - \nu\left(\frac{\sigma_{\alpha\beta\tau}(d)}{1-\lambda} + 1\right)\left[C_{\alpha\beta\tau} + \frac{C_1}{\sqrt{d}}\right]\kappa_\tau(d) \\ &=: r_0(\alpha,\beta,d) - \nu\, r_1(\alpha,\beta,\tau,\lambda,d) \\ &=: r_B(\alpha,\beta,\tau,\lambda,\nu,d)\,.\end{aligned} \tag{4.3.14}$$

With all the other conditions on the parameters imposed before, for sufficiently large $d$ we will have $r_0(\ldots) > 0$. Furthermore, we always have $r_1(\ldots) > 0$, so choosing $0 < \nu < r_0(\ldots)/r_1(\ldots)$ will guarantee $r_B(\ldots)$ to be positive.

**Step 4: Specification of $\mu$ and bounding of conditional probabilities.** We specify the measure $\mu$ on the set of functions $\{f_U\} \subset F^d_{\circ|}$ defined as in (4.3.4) with $U \subseteq W$. Remember that the $f(\mathbf{w})$ (for $\mathbf{w} \in W$) shall be independent Bernoulli random variables with probability $p = \mu\{f(\mathbf{w}) = 1\}$. Having the augmented information $\tilde{y} = (V_0, V_1)$, the values $f(\mathbf{w})$ are still independent random variables with probabilities

$$\mu_{\tilde{y}}\{f(\mathbf{w}) = 1\} = \begin{cases} 0 & \text{if } \mathbf{w} \in V_0, \\ 1 & \text{if } \mathbf{w} \in V_1, \\ p & \text{if } \mathbf{w} \in W \setminus (V_0 \cup V_1). \end{cases}$$

Then for $\mathbf{x} \in B$ we have the estimate

$$\mu_{\tilde{y}}\{f(\mathbf{x}) = 0\} \le (1-p)^M \le \exp\left(-p\lambda\binom{a}{t}\right) = \exp(-\lambda\varrho)\,,$$

where we write $p := \varrho/\binom{a}{t}$ with $0 < \varrho < \binom{a}{t}$. The other estimate is

$$\begin{aligned}\mu_{\tilde{y}}\{f(\mathbf{x}) = 0\} &\ge (1-p)^{\binom{b}{t}} = \exp\left(\log(1-p)\binom{b}{t}\right) \\ &\ge \exp\Bigg(-\varrho\,\sigma_{\alpha\beta\tau}(d)\underbrace{\left(\frac{1}{2} + \frac{1}{2\,(1 - \varrho/\gamma_{\alpha\tau}(d))}\right)}_{=:\kappa_{\varrho\gamma}(d)}\Bigg) =: q_0(\alpha,\beta,\tau,\varrho,d) \\ &\xrightarrow[d\to\infty]{} \exp\Big(-\varrho\,\exp\left((\beta-\alpha)\,\tau\right)\Big)\,,\end{aligned} \tag{4.3.15}$$

Here we used that, for $0 \le p < 1$,

$$0 \ge \log(1-p) = -\left(p + \sum_{k=2}^{\infty} \frac{p^k}{k}\right) \ge -\left(p + \sum_{k=2}^{\infty} \frac{p^k}{2}\right) = -p\left(\frac{1}{2} + \frac{1}{2\,(1-p)}\right),$$

together with the estimates

$$p\binom{b}{t} \le \varrho\, \sigma_{\alpha\beta\tau}(d)\,,$$

and, provided $t \le a$,

$$\binom{a}{t} \ge \left(\frac{a}{t}\right)^t \ge \left(\frac{\sqrt{d}+\alpha}{2\left(\tau + 1/\sqrt{d}\right)}\right)^{\tau\sqrt{d}} =: \gamma_{\alpha\tau}(d)\,.$$

Note that $\gamma_{\alpha\tau}(d) \xrightarrow[d\to\infty]{} \infty$ implies $\kappa_{\varrho\gamma}(d) \xrightarrow[d\to\infty]{} 1$. It follows that for $\mathbf{x} \in B$,

$$\begin{aligned}\min\Big\{\mu_{\tilde{y}}\{f(\mathbf{x})=1\},\, \mu_{\tilde{y}}\{f(\mathbf{x})=0\}\Big\}& \\ \ge \min\left\{1-\exp\left(-\varrho\,\lambda\right),\, q_0(\alpha,\beta,\tau,\varrho,d)\right\} &=: q(\alpha,\beta,\tau,\lambda,\varrho,d)\,. \quad (4.3.16)\end{aligned}$$

**Step 5: The final error bound.**
By Lemma 4.11 and Bakhvalov's technique (Proposition 1.2) we obtain the final estimate for $n \le \nu\, 2^{\tau\sqrt{d}} = \nu\, \exp(\sigma\sqrt{d})$, where $\sigma = \tau\, \log 2$,

$$\begin{aligned} e^{\mathrm{ran}}(n, \mathrm{APP}, F_{\circ|}^d, \Lambda^{\mathrm{std}}) &\ge e^{\mathrm{avg}}(n, \mathrm{APP}, \mu, \Lambda^{\mathrm{std}}) \\ [\text{any valid } \tilde{y}] \quad &\ge \frac{\#B}{\#\{0,1\}^d}\, \min\big\{\mu_{\tilde{y}}\{f(\mathbf{x})=0\},\, \mu_{\tilde{y}}\{f(\mathbf{x})=1\} \mid \mathbf{x}\in B\big\} \\ [(4.3.14) \text{ and } (4.3.16)] \quad &\ge [r_0(\alpha,\beta,\tau) - \nu\, r_1(\alpha,\beta,\tau,\lambda,d)] \cdot q(\alpha,\beta,\tau,\lambda,\varrho,d) \\ &=: \hat{\varepsilon}(\alpha,\beta,\tau,\lambda,\nu,\varrho,d)\,. \qquad (4.3.17)\end{aligned}$$

Fixing $d = d_0$, and with appropriate values for the other parameters, we can provide $r_B = r_0 - \nu\, r_1 > 0$. The value of $\varrho$ should be adapted for that $q(\ldots)$ is big (and positive in the first place). The function $\hat{\varepsilon}(\ldots, d)$ is monotonously increasing in $d$, so an error bound for $d = d_0$ implies error bounds for $d \ge d_0$ while keeping in particular $\nu$ and $\tau$. Clearly, for any $0 < \varepsilon_0 < \hat{\varepsilon}(\ldots)$, this gives lower bounds for the $\varepsilon$-complexity,

$$n^{\mathrm{ran}}(\varepsilon_0, \mathrm{APP}, F_{\circ|}^d, \Lambda^{\mathrm{std}}) > \nu\, \exp(\sigma\sqrt{d})\,.$$

Note that the definition of the measure does not depend on $n$. Moreover, by the above calculations, we have a general lower bound $\hat{\varepsilon}(n(\mathbf{y}))$ which holds for the conditional error in the case of varying cardinality as well. This estimate can be seen as a convex function in $\bar{n} \ge 0$, indeed, fixing all parameters but $\nu := \bar{n}\, 2^{-\tau\sqrt{d}}$, we have

$$\hat{\varepsilon}(\bar{n}) = \left[r_0 - \bar{n}\, 2^{-\tau\sqrt{d}}\, r_1\right]_+ q\,.$$

By Lemma 1.6 the lower bounds extend to methods with varying cardinality.

**Step 6: Smaller $\varepsilon$ and bigger exponent $\tau$ for higher dimensions.**[15] More sophisticated, if we have a lower bound $\hat{\varepsilon}(\alpha_0, \beta_0, \tau_0, \lambda, \nu, \varrho, d_0) > \varepsilon_0$, then for $\tau \geq \tau_0$ and $d \geq d_0 \left(\frac{\tau}{\tau_0}\right)^2$ we obtain the lower bound

$$\hat{\varepsilon}(\alpha(\tau), \beta(\tau), \tau, \lambda, \nu, \varrho, d) > \frac{\tau_0}{\tau}\, \varepsilon_0 =: \varepsilon$$

with $\alpha(\tau) = \alpha_0 \frac{\tau_0}{\tau}$ and $\beta(\tau) = \beta_0 \frac{\tau_0}{\tau}$, supposed that in addition we fulfil the conditions $\tau_0 \geq \beta_0$ and $-\tau_0 \leq \alpha_0 \leq 0$. This gives us a valid estimate

$$n^{\mathrm{ran}}(\varepsilon, \mathrm{APP}, F_{\circ|}^d, \Lambda^{\mathrm{std}}) \geq \nu\, 2^{\tau\sqrt{d}} = \nu\, 2^{\tau_0\, \varepsilon_0\, \sqrt{d}/\varepsilon}$$

under the restriction $d \geq d_0 \left(\frac{\varepsilon_0}{\varepsilon}\right)^2$.

In detail, the constraint $\tau \leq \tau_0 \sqrt{d/d_0}$ is needed to contain several correcting terms that occur because $a$, $b$, and $t$ can only take integer values. For example, from (4.3.8) we have the correcting factor $\kappa_\tau(d)$, for which holds

$$1 \leq \kappa_\tau(d) = 1/\sqrt{1 - \tau/\sqrt{d} - 1/d} \leq 1/\sqrt{1 - \tau_0/\sqrt{d_0} - 1/d_0} = \kappa_{\tau_0}(d_0)\,.$$

Furthermore, with the choice of $\alpha(\tau)$ and $\beta(\tau)$, the product $(\beta - \alpha)\tau = (\beta_0 - \alpha_0)\tau_0$ is kept constant. This is the key element for the estimate

$$\sigma_{\alpha\beta\tau}(d) \leq \sigma_{\alpha_0,\beta_0,\tau_0}(d_0)\,,$$

see its definition (4.3.13). For the $d$-dependent correcting terms that occur therein, we have $1 \leq \kappa_{\alpha\tau}(d) \leq \kappa_{\alpha_0\tau_0}(d_0)$, and $0 \leq K_{\alpha\beta\tau}(d) \leq K_{\alpha_0\beta_0\tau_0}(d_0)$, where the assumption $\alpha_0 \leq 0$ comes into play. For bounding $\kappa_{\alpha\tau}(d)$, we also need $\tau \leq \tau_0 \sqrt{d/d_0}$.

Having $\sigma_{\alpha\beta\tau}(d)$ under control, one can easily show[16]

$$q(\alpha(\tau), \beta(\tau), \tau, \lambda, \varrho, d) \geq q(\alpha_0, \beta_0, \tau_0, \lambda, \varrho, d_0)\,,$$

and, more complicated,

$$r_B(\alpha(\tau), \beta(\tau), \tau, \lambda, \nu, d) \geq \frac{\tau_0}{\tau}\, r_B(\alpha_0, \beta_0, \tau_0, \lambda, \nu, d_0).$$

For the latter we need in particular the inequalities

$$C_{\alpha\beta} \geq \frac{\tau_0}{\tau}\, C_{\alpha_0,\beta_0}\,, \quad \text{and} \quad C_{\alpha\beta\tau} \leq \frac{\tau_0}{\tau}\, C_{\alpha_0,\beta_0,\tau_0}\,.$$

We start with the first inequality,

$$\begin{aligned} C_{\alpha\beta} &= \frac{1}{\sqrt{2\pi}} \int_\alpha^\beta \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x \\ [x = \tfrac{\tau_0}{\tau}\, u] \quad &= \frac{\tau_0}{\tau\sqrt{2\pi}} \int_{\alpha_0}^{\beta_0} \underbrace{\exp\left(-\left(\frac{\tau_0}{\tau}\right)^2 \frac{u^2}{2}\right)}_{[\tau \geq \tau_0] \quad \geq \exp\left(-\frac{u^2}{2}\right)} \mathrm{d}u \\ &\geq \frac{\tau_0}{\tau}\, C_{\alpha_0,\beta_0}\,. \end{aligned}$$

[15] These considerations give results of a new quality compared to Blum et al. [8].

[16] This effectively means examining $q_0(\alpha(\tau), \beta(\tau), \tau, \varrho, d)$, see (4.3.15), where in particular we need to show $\gamma_{\alpha\tau}(d) \geq \gamma_{\alpha_0,\tau_0}(d_0)$, which relies on $\alpha_0 \leq 0$ and $\tau_0 \leq \tau \leq \tau_0 \sqrt{d/d_0}$.

The second inequality is a bit trickier,

$$\begin{aligned} C_{\alpha\beta\tau} &= \frac{1}{\sqrt{2\pi}} \int_{\alpha-\tau}^{\beta-\tau} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x \\ [\text{subst. } x+\tau = \tfrac{\tau_0}{\tau}(u+\tau_0)] \quad &= \frac{\tau_0}{\tau\sqrt{2\pi}} \int_{\alpha_0-\tau_0}^{\beta_0-\tau_0} \underbrace{\exp\left(-\frac{1}{2}\left(\frac{\tau_0}{\tau}(u+\tau_0)-\tau\right)^2\right)}_{\geq \exp\left(-\frac{u^2}{2}\right)} \mathrm{d}u \\ [\tfrac{\tau_0}{\tau}(u+\tau_0)-\tau \leq u \leq 0] \quad &\leq \frac{\tau_0}{\tau} C_{\alpha_0,\beta_0,\tau_0} \,. \end{aligned}$$

Here, $u \leq 0$ followed from the the upper integral boundary $u \leq \beta_0 - \tau_0$ and the assumption $\beta_0 \leq \tau_0$. The other constraint, $\psi(\tau) := \frac{\tau_0}{\tau}(u+\tau_0) - \tau \leq u$, followed from the monotonous decay of $\psi(\tau)$ for $\tau \geq \tau_0$, taking $\alpha_0 - \tau_0 \leq u$ from the lower integral boundary into account, and recalling the assumption $\alpha_0 \geq -\tau_0$.

**Step 7: Example for numerical values.**
The stated numerical values result from the setting $\alpha = -0.33794$, $\beta = 0.46332$, $\tau = 1.47566 > \frac{1}{\log 2}$ and $\lambda = 0.77399$. We adapt $\varrho = 0.25960$, and for starting dimension $d_0 = 100$ and $n_0 = 108$ (choosing $\nu$ appropriately) we obtain the lower error bound $\hat{\varepsilon}(\ldots) = 0.03333335... > \frac{1}{30} =: \varepsilon_0$. □

### 4.3.3 Remarks on the Proof

**Remark 4.13** (On finding parameters for good lower bounds)**.** Fixing $d_0$ and $n_0$, one may vary $\alpha$, $\beta$, $\tau$, and $\lambda$ so that the error bound $\hat{\varepsilon}(\ldots)$ is maximized (meanwhile adjusting $\nu = n_0\, 2^{-\tau\sqrt{d_0}}$ and $\varrho$). Numerical calculations indicate that $d_0 = 24$ is likely to be the first dimension where with $n_0 = 1$ we can obtain a positive error bound (which is at around $\varepsilon_0 \approx 10^{-6}$). Choosing $d_0 = n_0$ we first obtain a positive error bound $\varepsilon_0 \approx 3 \times 10^{-5}$ for $d_0 = 40$.

That way it is also possible to find the maximal value of $n_0$ such that for a given dimension $d_0$ the error bound exceeds a given value $\varepsilon_0$. In doing so, we find a result with a particular $\tau > 0$ and an estimate for the $\varepsilon_0$-complexity for $d \geq d_0$:

$$n^{\mathrm{ran}}(\varepsilon_0, d) \geq n_0\, 2^{\tau(\sqrt{d}-\sqrt{d_0})} = n_0 \exp(\sigma(\sqrt{d} - \sqrt{d_0}))\,.$$

The following tabular lists the maximal $n_0$ for given $d_0$ such that we still find a lower bound that exceeds $\varepsilon_0 = \frac{1}{30}$. In addition, we give the maximal possible value for $\tau$ (and $\sigma = \tau \log 2$) such that we still obtain the error bound $\varepsilon_0$ with the same $n_0$.

| $d_0$ | $n_0$ | $\tau$ | $\sigma = \tau \log 2$ |
|---|---|---|---|
| 51 | 1 | 1.0696 | 0.7414 |
| 100 | 108 | 1.4795 | 1.0255 |
| 200 | 498 098 | 1.9796 | 1.3721 |

As observable in the examples, the value for $\tau$ is increasing for growing dimension, so if we aim to find a good lower bound for the $\varepsilon_0$-complexity for a particular dimension $d$, it is preferable to use an estimate based on a big value $d_0 \leq d$. For example, for $d = 200$ we obtain

- $n^{\mathrm{ran}}(\varepsilon_0, d) > \quad 179$, based on $d_0 = 51$ and $\tau = 1.0696$,
- $n^{\mathrm{ran}}(\varepsilon_0, d) > \quad 7\,554$, based on $d_0 = 100$ and $\tau = 1.4795$,
- $n^{\mathrm{ran}}(\varepsilon_0, d) > 498\,098$, computed directly for $d_0 = 200$.

The question on how big the exponent can get is answered within the next remark.

**Remark 4.14** (Maximal value for the exponential constant $c = \sigma_0\,\varepsilon_0$)**.** We have results of the type

$$n^{\mathrm{ran}}(\varepsilon, d) \geq \nu \, \exp\left(c\,\frac{\sqrt{d}}{\varepsilon}\right)$$

that hold for "*large* $d$" and $\varepsilon \succeq 1/\sqrt{d}$. In the asymptotics of $d \to \infty$, any estimate with a larger exponent will outstrip an estimate with a smaller exponent, so it is preferable to have a big constant $c$ in the exponent, but $\nu$ can be arbitrarily small. In order to find the maximal value for $c$, we consider the limiting case for the detailed error bounds of Theorem 4.8, see Step 5 of the proof. First, we ask the question what error bounds are possible for a given $\tau$,

$$\lim_{\lambda\to 1}\lim_{\substack{d\to\infty\\ \nu\to 0}} \hat{\varepsilon}(\ldots) = (\Phi(\beta) - \Phi(\alpha)) \max_{\varrho>0} \min\left\{1 - \exp(-\varrho),\, \exp\Big(-\varrho\,\exp((\beta-\alpha)\,\tau)\Big)\right\}.$$

The maximal value for optimal $\alpha$, $\beta$, and $\varrho$ gives us a limiting value $\bar{\varepsilon}(\tau)$. Amongst all settings with constant difference $(\beta - \alpha)$, the factor $(\Phi(\beta) - \Phi(\alpha))$ is maximized (and hence also the asymptotic lower bound) for the symmetrical choice $\alpha = -\beta$. Writing $\beta\,\tau =: \vartheta$, we obtain

$$\bar{\varepsilon}(\tau) := \max_{\vartheta>0} \underbrace{\left(2\Phi\left(\frac{\vartheta}{\tau}\right) - 1\right) \max_{\varrho>0} \min\left\{1 - \exp(-\varrho),\, \exp\Big(-\varrho\,\exp(2\vartheta)\Big)\right\}}_{=:\tilde{\varepsilon}(\tau,\vartheta)},$$

so the second factor is formally independent from $\tau$ now.[17] The product $\tau\,\tilde{\varepsilon}(\tau,\vartheta)$ is growing with $\tau$ when $\vartheta$ is fixed. Therefore, the product $\tau\,\bar{\varepsilon}(\tau)$ is maximized in the limit $\tau \to \infty$,

$$\lim_{\tau\to\infty} \tau\,\bar{\varepsilon}(\tau) = \max_{\vartheta>0} \sqrt{\frac{2}{\pi}}\,\vartheta \max_{\varrho>0} \min\left\{1 - \exp(-\varrho),\, \exp\Big(-\varrho\,\exp(2\vartheta)\Big)\right\} \approx 0.1586\,.$$

In other words, switching the basis of the exponential expression, now considering $\sigma = \tau\,\log 2$, the constant $c = \sigma_0\,\varepsilon_0$ in Theorem 4.8 cannot exceed 0.1100 when relying on the given proof technique. For comparison, in the numerical example of the theorem with $d_0 = 100$, $n_0 = 108$, and $\varepsilon_0 = \frac{1}{30}$, we have $c = \frac{1}{30} \approx 0.3333$.

Compared to the upper bounds for Boolean functions, see Theorem 4.18 in Section 4.4, there is a significant gap in the exponent that can reach arbitrarily high factors if $1/\sqrt{d} \prec \varepsilon < 1/2$ (the growth, however, is only logarithmic).

[17] Compare with the choice of $\alpha(\tau)$ and $\beta(\tau)$ in Step 5 within the proof of Theorem 4.8.

**Remark 4.15** (Close to the initial error)**.** We discuss necessary modifications to the proof of Theorem 4.8 in order to reproduce the result of Blum, Burch, and Langford [8, Sec 4], which states that for sufficiently large dimensions $d$, and $n \geq d$, we have

$$e^{\mathrm{ran}}(n, \mathrm{APP}, F^d_{\circ|}, \Lambda^{\mathrm{std}}) \geq \frac{1}{2} - C\,\frac{\log(d\,n)}{\sqrt{d}}\,, \tag{4.3.18}$$

where $C > 0$ is a numerical constant.

We start with direct modifications for a weaker version of (4.3.18). Several parameters are chosen with regard to the estimates of Step 3. First, taking $-\alpha = \beta = \sqrt{(\log d)/2}$, by Hoeffding bounds we have

$$\frac{\#D_{ab}}{\#\{0,1\}^d} \geq 1 - \frac{2}{\sqrt{d}}\,,$$

compare Step 3.1. The calculations of Step 3.2 can be replaced by the trivial estimate

$$\frac{\#(Q_{\mathbf{w}} \cap D_{ab})}{\#Q_{\mathbf{w}}} \leq 1\,.$$

We choose $\nu = 1/d$ and $\lambda = 1 - 1/\sqrt{d}$, thus the final estimate (4.3.14) of Step 3.4 reduces to

$$\frac{\#B}{\#\{0,1\}^d} \geq 1 - \frac{1}{d} - \frac{2 + \sigma_{\alpha\beta\tau}(d)}{\sqrt{d}}\,. \tag{4.3.19}$$

The choice of $\nu$ implies that, for given $n$, we need to take

$$\tau := \frac{\log_2(d\,n)}{\sqrt{d}}\,,$$

thus $t := \lceil \log_2(d\,n) \rceil$. Note that $\sigma_{\alpha\beta\tau}(d)$ can be bounded by a constant as long as $\beta\tau \preceq 1$, which is equivalent to $n \leq \exp(c\,\sqrt{d/\log d})$ for some $c > 0$. This log-term in the exponent is unpleasant when trying to reproduce the result of Blum et al. The term $\sigma_{\alpha\beta\tau}(d)$ occurs from estimating $\#V_0$, see (4.3.6) in Step 2. In the original paper, for this purpose, Chernoff bounds are used, and we do not need to include the boundary value $b$ in the definition of the measure, see Step 1. Then the cardinality of $\#B$ can be estimated by terms that only depend on $\tau$ and $d$, the term $\sigma_{\alpha\beta\tau}(d)$ can be replaced by a constant.

More effort than before has to be put in estimating the conditional distribution of $f(\mathbf{x})$ for $\mathbf{x} \in B$, compare Step 4. For detailed calculations refer to the original proof of Blum et al. [8]. The parameter $p$ of the distribution is determined by the equation

$$(1-p)^{\binom{d/2}{t}} \stackrel{!}{=} \frac{1}{2}\,,$$

thus, for $|\mathbf{x}|_1 = \frac{d}{2}$, the function values $f(\mathbf{x})$ are 0 and 1 with equal probability under $\mu$. Then we obtain

$$\begin{aligned} \mu_{\tilde{y}}\{f(\mathbf{x}) = 0\} &\leq 2^{\lambda\binom{a}{t}/\binom{d/2}{t}} \leq \frac{1}{2} + \frac{(\beta+1)\,t}{\sqrt{d}}\,, \quad \text{and} \\ \mu_{\tilde{y}}\{f(\mathbf{x}) = 0\} &= 2^{-\binom{b}{t}/\binom{d/2}{t}} \geq \frac{1}{2} - \frac{(\beta+1)\,t}{\sqrt{d}}\,. \end{aligned} \tag{4.3.20}$$

Note that these two inequalities become trivial if $(\beta+1)\,t/\sqrt{d}$ exceeds $\frac{1}{2}$. In particular, $(\beta+1)\,t/\sqrt{d} \le \frac{1}{2}$ is an assumption needed for the proof of the second inequality when following the steps in Blum et al.

Combining (4.3.19) and (4.3.20), and inserting the values for $\beta$ and $t$, we obtain the estimate

$$e(n,d) \ge \frac{1}{2} - C\,\frac{(1+\sqrt{\log d})\,\log d\,n}{\sqrt{d}}\,,$$

with $C>0$ being a numerical constant. This is a weaker version of the result (4.3.18) by a logarithmic factor in $d$.

Blum et al. proved a stronger version without this logarithmic factor by integrating over $\beta$ from 1 to $\sqrt{d}$. The weaker version with constant $\beta = \sqrt{(\log d)/2}$ has also been mentioned in the original paper already. The integration over $\beta$ is only possible if we estimate $\#V_1$ by Chernoff bounds, the boundary value $b$ may not be part of the definition of the functions for the measure $\mu$. Interestingly, this integral runs also over such $\beta$ where the estimates on the conditional measure (4.3.20) give negative values, thereby weakening the lower bounds. Still, this refined proof technique gives an improvement by a logarithmic term.

**Remark 4.16** (A combined lower bound for real-valued monotone functions)**.** Similarly to Remark 4.7, which was about the deterministic setting, we can find lower bounds for the Monte Carlo approximation of real-valued monotone functions that include arbitrarily small $\varepsilon$ for small dimensions already, but still reflect the $d$-dependency known from Theorem 4.8.

With the notation from Remark 4.7, given $\mathbf{m}\in\mathbb{N}^d$, we split the domain into $\prod\mathbf{m}$ sub-cuboids $C_{\mathbf{i}}$, and consider monotone functions $f\in F^d_{\mathrm{mon}}$ that on each cuboid only have function values within an interval of length $1/(|\mathbf{m}|_1-d+1)$ such that monotonicity is guaranteed whenever the function is monotone on each of the sub-cuboids. Having lower bounds from Theorem 4.8,

$$e^{\mathrm{ran}}(n=\nu\,2^{\tau\sqrt{d}},\mathrm{APP},F^d_{\mathrm{mon}},\Lambda^{\mathrm{std}}) \ge (r_0-\nu\,r_1)\,q =: \varepsilon_1\,,$$

see also (4.3.17) for the inner structure of the lower bound, we can estimate the error we make on each of the sub-cuboids using $n_{\mathbf{i}}=\nu_{\mathbf{i}}\,2^{\tau\sqrt{d}}$ function values only,

$$\begin{aligned} e^{\mathrm{ran}}(n=\nu\,2^{\tau\sqrt{d}},\mathrm{APP},F^d_{\mathrm{mon}},\Lambda^{\mathrm{std}}) &\ge \frac{1}{|\mathbf{m}|_1-d+1}\,\left[\frac{1}{\prod\mathbf{m}}\sum_{\mathbf{i}}(r_0-\nu_{\mathbf{i}}\,r_1)\,q\right] \\ &= \frac{1}{|\mathbf{m}|_1-d+1}\,\left(r_0-\frac{\nu}{\prod\mathbf{m}}\right)\,q\,, \end{aligned}$$

where $n=\sum_{\mathbf{i}} n_{\mathbf{i}}$. For $0<\varepsilon<\varepsilon_1$, choose an appropriate splitting parameter $\mathbf{m}$, and obtain the complexity bound

$$n^{\mathrm{det}}(\varepsilon,\mathrm{APP},F^d_{\mathrm{mon}},\Lambda^{\mathrm{std}}) \ge \begin{cases} \nu\,2^{\tau\sqrt{d}+\lfloor\varepsilon_1/\varepsilon\rfloor-2} & \text{for } \varepsilon\in[\frac{\varepsilon_1}{d+1},\varepsilon_1], \\ \nu\,2^{\tau\sqrt{d}+d\,\log_2\lfloor\varepsilon_1/(\varepsilon\,d)\rfloor)-1} & \text{for } \varepsilon\in(0,\frac{\varepsilon_1}{d+1}]. \end{cases} \tag{4.3.21}$$

Knowing

$$n^{\mathrm{ran}}(\varepsilon, \mathrm{APP}, F^d_{\circ|}, \Lambda^{\mathrm{std}}) > \nu\, 2^{c\sqrt{d}/\varepsilon}\,,$$

for $d > d_0$ and $\varepsilon \geq \varepsilon_0\sqrt{d_0/d}$, we can take (4.3.21) with the $d$-dependent values $\varepsilon_1 = \varepsilon_0\sqrt{d_0/d}$ and $\tau = c/\varepsilon_1 = c/\varepsilon_0\sqrt{d/d_0}$ in order to get enhanced error bounds for $0 < \varepsilon < \varepsilon_0\sqrt{d_0/d}$.

## 4.4 Breaking the Curse with Monte Carlo

A new algorithm for the approximation of real-valued monotone functions on the unit cube is presented and analysed in Section 4.4.2. It is the first algorithm to show that for this problem the curse of dimensionality does not hold in the randomized setting. The idea is directly inspired by a method for Boolean monotone functions due to Bshouty and Tamon [9]. We start with the presentation of the less complicated Boolean case in Section 4.4.1. The structure of the proofs in each of the two sections is analogous to the greatest possible extend so that one can always find the counterpart within the other setting (if there is one).

### 4.4.1 A Known Method for Boolean Functions

We present a method known from Bshouty and Tamon [9] for the randomized approximation of Boolean functions that comes close to the lower bounds from Theorem 4.8. Actually, they considered a slightly more general setting, allowing product weights for the importance of different entries of a Boolean function $f$, we only study the special case that fits to our setting. The analysis of the original paper was done for the *margin of error* setting, but it can be easily converted into results on the Monte Carlo error as we prefer to define it by means of expectation.[18]

For the formulation and the analysis of the algorithm it is convenient to redefine the notion of Boolean functions and to consider the class of functions

$$G^d_{\pm} := \{f : \{-1,+1\}^d \to \{-1,+1\}\}\,,$$

the input set of Boolean functions is renamed $F^d_{\pm} \subset G^d_{\pm}$. We keep the distance dist that we had for the old version of $G^d_{\circ|}$, for $f_1, f_2 \in G^d_{\pm}$ we have

$$\mathrm{dist}(f_1, f_2) := \frac{1}{2^d}\,\#\{\mathbf{i} \in \{-1,1\}^d \mid f_1(\mathbf{i}) \neq f_2(\mathbf{i})\}\,,$$

compare (4.1.3), thus the diameter of $G^d_{\pm}$ is still 1. This metric differs by a factor 2 from the induced metric that we obtain when regarding $G^d_{\pm}$ as a subset of the

[18] In the margin of error setting we want to determine $\varepsilon, \delta > 0$ such that for any input function $f$ the actual error of the randomized algorithm exceeds $\varepsilon$ only with probability $\delta$. Since for Boolean functions the error cannot exceed 1, the corresponding expected error is bounded from above by $\varepsilon + \delta$.
Conversely, for any $\varepsilon$ that exceeds the Monte Carlo error $e^{\mathrm{ran}}$, we obtain $\delta \leq e^{\mathrm{ran}}/\varepsilon$ for the uncertainty level by Chebyshev's inequality. Practically, if we aim for a small $\delta$, we lose a lot in this direction, and it is advisable to analyse the margin of error setting directly whenever it is of interest.

Euclidean space $L_2(\mathrm{unif}\{-1,+1\}^d)$. For this space we choose the orthonormal basis $\{\psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\{0,1\}^d}$,

$$\psi_{\boldsymbol{\alpha}}(\mathbf{x}) := \mathbf{x}^{\boldsymbol{\alpha}} = \prod_{j=1}^{d} x_j^{\alpha_j} \quad \text{for } \mathbf{x} \in \{-1,+1\}^d.$$

Every Boolean function can be written as the Fourier decomposition

$$f = \sum_{\boldsymbol{\alpha}\in\{0,1\}^d} \hat{f}(\boldsymbol{\alpha})\, \psi_{\boldsymbol{\alpha}}$$

with the Fourier coefficients

$$\hat{f}(\boldsymbol{\alpha}) := \langle \psi_{\boldsymbol{\alpha}}, f\rangle = \mathbb{E}\, \mathbf{X}^{\boldsymbol{\alpha}}\, f(\mathbf{X})\,,$$

where $\mathbf{X}$ is uniformly distributed on $\{-1,+1\}^d$. The idea of the algorithm is to use random samples $f(\mathbf{X}_1),\ldots,f(\mathbf{X}_n)$, with $\mathbf{X}_i \overset{\text{iid}}{\sim} \mathrm{unif}\{-1,+1\}^d$, in order to approximate the low-degree Fourier coefficients for $|\boldsymbol{\alpha}|_1 \le k$, $k \in \mathbb{N}$,

$$\hat{f}(\boldsymbol{\alpha}) \approx \hat{h}(\boldsymbol{\alpha}) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^{\boldsymbol{\alpha}}\, f(\mathbf{X}_i)\,.$$

Based on the Fourier approximation

$$h := \sum_{\substack{\boldsymbol{\alpha}\in\{0,1\}^d \\ |\boldsymbol{\alpha}|_1\le k}} \hat{h}(\boldsymbol{\alpha})\, \psi_{\boldsymbol{\alpha}}\,,$$

we return the output $g := A^{\omega}_{n,k}(f)$ with

$$g(\mathbf{x}) := \operatorname{sgn} h(\mathbf{x}) = \begin{cases} +1 & \text{if } h(\mathbf{x}) \ge 0, \\ -1 & \text{if } h(\mathbf{x}) < 0. \end{cases}$$

We will give a complete analysis of the above algorithm which is based on $L_2$-approximation. In fact, from

$$f(\mathbf{x}) \ne g(\mathbf{x}) \quad \Leftrightarrow \quad f(\mathbf{x}) \ne \operatorname{sgn} h(\mathbf{x}) \quad \Rightarrow \quad (f(\mathbf{x}) - h(\mathbf{x}))^2 \ge 1$$

we obtain

$$\mathrm{dist}(f,g) \le \|f-h\|_{L_2}^2\,. \tag{4.4.1}$$

Note that every Boolean function $f : \{-1,+1\}^d \to \{-1,+1\}$ has norm 1 in the $L_2$-norm, so by Parseval's equation,

$$\sum_{\boldsymbol{\alpha}\in\{0,1\}^d} \hat{f}^2(\boldsymbol{\alpha}) = \|f\|_{L_2} = 1\,.$$

A key result for the analysis of the above algorithm is the following fact about the Fourier coefficients that are dropped.

**Lemma 4.17** (Bshouty and Tamon [9, Sec 4]). *For any monotone Boolean function $f$ we have*

$$\sum_{\substack{\boldsymbol{\alpha}\in\{0,1\}^d \\ |\boldsymbol{\alpha}|_1>k}} \hat{f}^2(\boldsymbol{\alpha}) \le \frac{\sqrt{d}}{k+1}\,.$$

*Proof.* Within the first step, we consider the special Fourier coefficients $\hat{f}(\mathbf{e}_j)$, which measure the sensitivity of $f$ with respect to a single variable $x_j$. These are the only Fourier coefficients where monotonicity guarantees a non-negative value. For $j = 1, \ldots, d$ we consider the restricted functions

$$\begin{aligned} f_{-j}(\mathbf{x}) &= f(\mathbf{z})\,, \quad \text{with } z_{j'} = x_{j'} \text{ for } j' \neq j, \text{ and } z_j = -1, \\ f_{+j}(\mathbf{x}) &= f(\mathbf{z})\,, \quad \text{with } z_{j'} = x_{j'} \text{ for } j' \neq j, \text{ and } z_j = +1. \end{aligned}$$

Due to the monotonicity of $f$, we have $f_{-j} \le f_{+j}$, using this and Parseval's equation, we obtain

$$\begin{aligned} \hat{f}(\mathbf{e}_j) = <\psi_{\mathbf{e}_j}, f> &= \frac{1}{2^d} \sum_{\mathbf{x}\in\{-1,+1\}^d} x_j\, f(\mathbf{x}) \\ &= \frac{1}{2^d} \sum_{\mathbf{x}\in\{-1,+1\}^d} \underbrace{\frac{f_{+j}(\mathbf{x}) - f_{-j}(\mathbf{x})}{2}}_{\in\{0,+1\}} \\ &= \left\| \frac{f_{+j} - f_{-j}}{2} \right\|_{L_2}^2 \\ &= \sum_{\boldsymbol{\alpha}\in\{0,1\}^d} \left\langle \psi_{\boldsymbol{\alpha}}, \frac{f_{+j} - f_{-j}}{2} \right\rangle^2 . \end{aligned}$$

Since the functions $f_{-j}$ and $f_{+j}$ are independent from $x_j$, the summands with $\alpha_j = 1$ vanish. For all the other summands, with $\alpha_j = 0$, we have

$$\begin{aligned} \left\langle \psi_{\boldsymbol{\alpha}}, \frac{f_{+j} - f_{-j}}{2} \right\rangle &= \frac{1}{2^d} \sum_{\mathbf{x}\in\{-1,+1\}^d} \mathbf{x}^{\boldsymbol{\alpha}}\, \frac{f_{+j}(\mathbf{x}) - f_{-j}(\mathbf{x})}{2} \\ &= \frac{1}{2^d} \sum_{\mathbf{x}\in\{-1,+1\}^d} \underbrace{x_j\, \mathbf{x}^{\boldsymbol{\alpha}}}_{=\mathbf{x}^{\boldsymbol{\alpha}'}} f(\mathbf{x}) \\ &= \langle \psi_{\boldsymbol{\alpha}'}, f\rangle = \hat{f}(\boldsymbol{\alpha}')\,, \end{aligned}$$

where $\alpha'_{j'} = \alpha_{j'}$ for $j' \neq j$, and $\alpha'_j = 1$. This leads to the identity

$$\hat{f}(\mathbf{e}_j) = \sum_{\substack{\boldsymbol{\alpha}\in\{0,1\}^d \\ \alpha_j=1}} \hat{f}^2(\boldsymbol{\alpha})\,.$$

Summing up over all dimensions, we obtain

$$\sum_{j=1}^{d} \hat{f}(\mathbf{e}_j) = \sum_{j=1}^{d} \sum_{\substack{\boldsymbol{\alpha}\in\{0,1\}^d \\ \alpha_j=1}} \hat{f}^2(\boldsymbol{\alpha}) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^d} |\boldsymbol{\alpha}|_1\, \hat{f}^2(\boldsymbol{\alpha}) \ge (k+1) \sum_{\substack{\boldsymbol{\alpha}\in\{0,1\}^d \\ |\boldsymbol{\alpha}|_1>k}} \hat{f}^2(\boldsymbol{\alpha})\,.$$

Finally,

$$1 = \sum_{\boldsymbol{\alpha}\in\{0,1\}^d} \hat{f}^2(\boldsymbol{\alpha}) \geq \sum_{j=1}^{d} \hat{f}^2(\mathbf{e}_j) \geq \frac{1}{d}\left(\sum_{j=1}^{d} \hat{f}(\mathbf{e}_j)\right)^2 ,$$

which, combined with the inequality above, proves the lemma. □

This helps us to obtain the following error and complexity bound, which is a simplification of Bshouty and Tamon [9, Thm 5.1], where the setting was more general, and the more demanding margin of error was considered.

**Theorem 4.18.** *For the algorithm $A_{n,k} = (A^{\omega}_{n,k})_{\omega\in\Omega}$, $k \in \{1,\dots,d\}$, we have the error bound*

$$e(A_{n,k}, F^d_{\pm} \hookrightarrow G^d_{\pm}) \leq \frac{\sqrt{d}}{k+1} + \frac{\exp\left(k\,\left(1+\log\frac{d}{k}\right)\right)}{n} .$$

*In particular, given $0 < \varepsilon < \frac{1}{2}$, the $\varepsilon$-complexity of the Monte Carlo approximation of monotone Boolean functions is bounded by*

$$n^{\mathrm{ran}}(\varepsilon, F^d_{\pm} \hookrightarrow G^d_{\pm}, \Lambda^{\mathrm{std}}) \leq \min\left\{\exp\left(C\,\frac{\sqrt{d}}{\varepsilon}\,\left(1+[\log\sqrt{d}\,\varepsilon]_+\right)\right),\, 2^d\right\} ,$$

*where $C > 0$ is some constant. Hence the curse of dimensionality does* not *hold.*

*Proof.* We first compute the accuracy at which we approximate each of the Fourier coefficients,

$$\begin{aligned}
\mathbb{E}[\hat{f}(\boldsymbol{\alpha}) - \hat{h}(\boldsymbol{\alpha})]^2 &= \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i=1}^{n}(\hat{f}(\boldsymbol{\alpha}) - \psi_{\boldsymbol{\alpha}}(\mathbf{X}_i)\,f(\mathbf{X}_i))\right]^2 \\
\text{[independent mean 0 variables]} \quad &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[\hat{f}(\boldsymbol{\alpha}) - \underbrace{\psi_{\boldsymbol{\alpha}}(\mathbf{X}_i)\,f(\mathbf{X}_i)}_{\in\{-1,+1\}}]^2 \qquad (4.4.2) \\
&\leq \frac{1}{n} .
\end{aligned}$$

This estimate is needed only for the set $A := \{\boldsymbol{\alpha} \in \{0,1\}^d \mid |\boldsymbol{\alpha}|_1 \leq k\}$, we obtain the general bound

$$\begin{aligned}
\mathbb{E}\,\mathrm{dist}(f,g) &\overset{(4.4.1)}{\leq} \mathbb{E}\,\|f-h\|^2_{L_2} \\
&\leq \sum_{\substack{\boldsymbol{\alpha}\in\{0,1\}^d \\ |\boldsymbol{\alpha}|_1 > k}} \hat{f}^2(\boldsymbol{\alpha}) + \sum_{\substack{\boldsymbol{\alpha}\in\{0,1\}^d \\ |\boldsymbol{\alpha}|_1 \leq k}} \mathbb{E}[\hat{f}(\boldsymbol{\alpha}) - \hat{h}(\boldsymbol{\alpha})]^2 \\
\text{[Lem 4.17 and (4.4.2)]} \quad &\leq \frac{\sqrt{d}}{k+1} + \frac{\#A}{n} .
\end{aligned}$$

By Lemma B.1, for $k \in \{1,\dots,d\}$, we estimate

$$\#A = \sum_{l=0}^{k}\binom{d}{l} \leq \left(\frac{\mathrm{e}\,d}{k}\right)^k = \exp\left(k\,\left(1+\log\frac{d}{k}\right)\right) .$$

This gives us the error bound for the Monte Carlo method $A_{n,k}$.

Choosing $k := \lfloor 2\sqrt{d}/\varepsilon \rfloor$ guarantees $\sqrt{d}/(k+1) \le \varepsilon/2$. The second term can be bounded by $\varepsilon/2$ if we choose

$$n := \left\lceil \frac{2\,\#A}{\varepsilon} \right\rceil \le \left\lceil \frac{2}{\varepsilon} \exp\left( \frac{2\sqrt{d}}{\varepsilon} \left(1 + [\log \sqrt{d}\,\varepsilon]_+\right)\right)\right\rceil .$$

For $\varepsilon \le 2/\sqrt{d}$, however, according to the error estimate of $A_{n,k}$, we would need to take $k = d$. In this case $\#A = 2^d$, and $n$ should be even larger. But then deterministically collected complete information $n = 2^d$ is the best solution with already exact approximation. □

**Remark 4.19** (Limiting case)**.** The present upper bounds fit the lower bounds from Theorem 4.8 up to a factor in the exponent which is logarithmic in $\varepsilon$ and $d$, but if we consider a sequence $\varepsilon_d = \varepsilon_0/\sqrt{d}$, there is no logarithmic gap at all.

There is a natural transition to complete information as $n$ approaches $2^d$. Indeed, take $n = 2^k$ for some natural number $k \le d$ and sample $f$ from function values computed for independent $\mathbf{X}_i$ chosen uniformly from $2^k$ disjoint subsets in $\{-1,+1\}^d$ of equal size. For instance, let the first $k$ entries within the random vector $\mathbf{X}_i$ be given by the binary representation of $i$, and let the remaining $d-k$ entries be independent Bernoulli random variables. The calculation (4.4.2) will still work essentially the same.

**Remark 4.20** (Non-interpolatory)**.** The algorithm $A_{n,k}$ is not always consistent with the knowledge we actually have on the function, and it does not even preserve monotonicity in general, so it is *non-interpolatory*.

Take, for example, $d \ge 2$ and $k = 1$, and the constant function $f = 1$. Assume that – for bad luck – all the sample points $\mathbf{X}_i$ happened to be $(-1,\dots,-1)$. Of course, the information $f(-1,\dots,-1) = 1$ already implies $f = 1$, thanks to monotonicity. But with the Fourier based algorithm,

$$h(\mathbf{x}) = 1 - \sum_{i=1}^{d} x_i$$

and for the output we have

$$g(1,\dots,1) = -1 < g(-1,\dots,-1) = +1\,,$$

which violates monotonicity.

We could modify the output, making the algorithm interpolatory, actually this is possible without affecting the error bounds. Obviously, it is an improvement to replace the original output $g$ by

$$g'(\mathbf{x}) := \begin{cases} +1 & \text{if } \exists i:\ \mathbf{X}_i \le \mathbf{x} \text{ and } f(\mathbf{X}_i) = +1, \\ -1 & \text{if } \exists i:\ \mathbf{X}_i \ge \mathbf{x} \text{ and } f(\mathbf{X}_i) = -1, \\ g(\mathbf{x}) & \text{else.} \end{cases}$$

Restoring monotonicity for the output is a bit more complicated since one needs to survey the output $g$ as a whole.[19] The general idea is to find pairs of points $\mathbf{z}_1 \leq \mathbf{z}_2$ with $g(\mathbf{z}_1) = +1 > g(\mathbf{z}_2) = -1$. At least one of these values is a misprediction of the input $f$. If we flip these values, that is, we create a new output

$$g'(\mathbf{x}) := \begin{cases} -1 & \text{for } \mathbf{x} = \mathbf{z}_1, \\ +1 & \text{for } \mathbf{x} = \mathbf{z}_2, \\ g(\mathbf{x}) & \text{else,} \end{cases}$$

then at least one of these values predicts $f$ correctly, so $g'$ approximates $f$ not worse than $g$ does. We could proceed like this until we obtain an interpolatory output, if needed.

### 4.4.2 Real-Valued Monotone Functions

We present a generalization of the method from the above section to the situation of real-valued monotone functions. The method is based on Haar wavelets. For convenience, we change the range and now consider monotone functions

$$f : [0,1]^d \to [-1,+1]\,,$$

the altered input set shall be named $F^d_{\text{mon}\pm}$.[20]

We define dyadic cuboids on $[0,1]^d$ indexed by $\boldsymbol{\alpha} \in \mathbb{N}^d$, or equivalently by an index vector pair $(\boldsymbol{\lambda}, \boldsymbol{\kappa})$ with $\boldsymbol{\lambda} \in \mathbb{N}_0^d$ and $\boldsymbol{\kappa} \in \mathbb{N}_0^d$, $\kappa_j < 2^{\lambda_j}$, such that $\alpha_j \equiv 2^{\lambda_j} + \kappa_j$ for $j = 1, \ldots, d$:

$$C_{\boldsymbol{\alpha}} = C_{\boldsymbol{\lambda},\boldsymbol{\kappa}} := \bigtimes_{j=1}^{d} I_{\alpha_j}\,,$$

where

$$I_{\alpha_j} = I_{\lambda_j,\kappa_j} := \begin{cases} [\kappa_j\, 2^{-\lambda_j}, (\kappa_j + 1)\, 2^{-\lambda_j}) & \text{for } \kappa_j = 0, \ldots, 2^{\lambda_j} - 2, \\ [1 - 2^{-\lambda_j}, 1] & \text{for } \kappa_j = 2^{\lambda_j} - 1. \end{cases}$$

Note that for fixed $\lambda_j$ we have a decomposition of the unit interval $[0,1]$ into $2^{\lambda_j}$ disjoint intervals of length $2^{-\lambda_j}$. One-dimensional Haar wavelets $h_{\alpha_j} : [0,1] \to \mathbb{R}$ are defined for $\alpha_j \in \mathbb{N}_0$ (if $\alpha_j = 0$, we set $\lambda_j = -\infty$ and $\kappa_j = 0$),

$$h_{\alpha_j} := \begin{cases} \mathbb{1}_{[0,1]} & \text{if } \alpha_j = 0 \text{ (i.e. } \lambda_j = -\infty \text{ and } k = 0), \\ 2^{\lambda_j/2}\,(\mathbb{1}_{I_{\lambda_j+1,2\kappa_j+1}} - \mathbb{1}_{I_{\lambda_j+1,2\kappa_j}}) & \text{if } \alpha_j \geq 1 \text{ (i.e. } \lambda_j \geq 0). \end{cases}$$

[19] Usually we would not store all values of an approximant $g$ in a computer but only the coefficients that are necessary for a computation of $g(\mathbf{x})$ on demand. This process should be significantly cheaper than asking the oracle for a value $f(\mathbf{x})$, compare Remark 4.24.

[20] By the bijection $F^d_{\text{mon}} = \frac{1}{2}\,(F^d_{\text{mon}\pm} + 1)$, we can transfer results for $F^d_{\text{mon}\pm}$ to results for $F^d_{\text{mon}}$, which comes along with a reduction of the error quantities by a factor $\frac{1}{2}$.
For proofs on lower bounds it was more convenient to have functions $f : [0,1]^d \to [0,1]$, because then the distance of Boolean functions coincides with the $L_1$-distance of the corresponding subcubewise constant functions.

In $L_2([0,1]^d)$ we have the orthonormal basis $\{\psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}_0^d}$ with

$$\psi_{\boldsymbol{\alpha}}(\mathbf{x}) := \prod_{j=1}^{d} h_{\alpha_j}(x_j)\,.$$

The volume of the support of $\psi_{\boldsymbol{\alpha}}$ is $2^{-|\boldsymbol{\lambda}|_+}$ with $|\boldsymbol{\lambda}|_+ := \sum_{j=1}^d \max\{0,\lambda_j\}$. The basis function $\psi_{\boldsymbol{\alpha}}$ only takes discrete values $\{0, \pm 2^{|\boldsymbol{\lambda}|_+/2}\}$, hence it is normalized indeed.

We can write any monotone function $f$ as the Haar wavelet decomposition

$$f = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^d} \tilde{f}(\boldsymbol{\alpha})\,\psi_{\boldsymbol{\alpha}}$$

with the wavelet coefficients

$$\tilde{f}(\boldsymbol{\alpha}) := \langle \psi_{\boldsymbol{\alpha}}, f\rangle = \mathbb{E}\,\psi_{\boldsymbol{\alpha}}(\mathbf{X})\,f(\mathbf{X})\,,$$

where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$. For the algorithm we will use random samples $f(X_1),\dots,f(X_n)$, with $\mathbf{X}_i \overset{\text{iid}}{\sim} \text{unif}[0,1]^d$, in order to approximate the most important wavelet coefficients

$$\tilde{f}(\boldsymbol{\alpha}) \approx \tilde{g}(\boldsymbol{\alpha}) := \frac{1}{n}\sum_{i=1}^{n} \psi_{\boldsymbol{\alpha}}(\mathbf{X}_i)\,f(\mathbf{X}_i)\,.$$

In particular, we choose a resolution $r\in\mathbb{N}$, and a parameter $k\in\{1,\dots,d\}$, and only consider indices $\boldsymbol{\alpha}\equiv(\boldsymbol{\lambda},\boldsymbol{\kappa})$ with $\lambda_j < r$ and $|\boldsymbol{\alpha}|_0 := \#\{j \mid \alpha_j > 0\} < k$. The Monte Carlo method $(A^{\omega}_{n,k,r})_{\omega\in\Omega}$ will give the output

$$g := A^{\omega}_{n,k,r}(f) := \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d\\ |\boldsymbol{\alpha}|_0\le k\\ \boldsymbol{\lambda}<r}} \tilde{g}(\boldsymbol{\alpha})\,\psi_{\boldsymbol{\alpha}}\,.$$

We start with an analogue of Lemma 4.17.

**Lemma 4.21.** *For any monotone function $f\in F^d_{\text{mon}\pm}$ we have*

$$\sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d\\ |\boldsymbol{\alpha}|_0> k\\ \boldsymbol{\lambda}<r}} \tilde{f}(\boldsymbol{\alpha})^2 \le \frac{\sqrt{d\,r}}{k+1}\,.$$

*Proof.* Within the first step, we consider special wavelet coefficients $\tilde{f}(\alpha\,\mathbf{e}_j)$ that measure the average growth of $f$ for the $j$-th coordinate within the interval $I_\alpha$. We will frequently use the alternative indexing $I_{\lambda,\kappa}$ with $\alpha = 2^\lambda + \kappa \in \mathbb{N}$, where $\lambda\in\mathbb{N}_0$ and $\kappa = 0,\dots,2^\lambda - 1$. We define the two functions

$$f_{-\alpha j}(\mathbf{x}) := \begin{cases} 0 & \text{if } x_j \notin I_\alpha,\\ 2^{\lambda+1}\int_{I_{(\lambda+1,2\kappa)}} f(\mathbf{z})\Big|_{\substack{z_{j'}=x_{j'}\\ \text{for } j\neq j'}} \mathrm{d}z_j & \text{if } x_j\in I_\alpha,\end{cases}$$

$$f_{+\alpha j}(\mathbf{x}) := \begin{cases} 0 & \text{if } x_j \notin I_\alpha,\\ 2^{\lambda+1}\int_{I_{(\lambda+1,2\kappa+1)}} f(\mathbf{z})\Big|_{\substack{z_{j'}=x_{j'}\\ \text{for } j\neq j'}} \mathrm{d}z_j & \text{if } x_j\in I_\alpha.\end{cases}$$

Due to monotonicity of $f$, we have $f_{-\alpha j} \leq f_{+\alpha j}$. Using this and Parseval's equation, we obtain

$$\begin{aligned}
\tilde{f}(\alpha\,\mathbf{e}_j) = \langle \psi_{\alpha\,\mathbf{e}_j}, f\rangle &= 2^{\lambda/2}\,\left[\langle \mathbb{1}_{I_{\lambda+1,2\kappa+1}}, f\rangle - \langle \mathbb{1}_{I_{\lambda+1,2\kappa}}, f\rangle\right] \\
&= 2^{\lambda/2}\left\|\underbrace{\frac{f_{+\alpha j} - f_{-\alpha j}}{2}}_{\in[0,1]}\right\|_{L_1} \\
&\geq 2^{\lambda/2}\left\|\frac{f_{+\alpha j} - f_{-\alpha j}}{2}\right\|_{L_2}^2 \\
&= 2^{\lambda/2}\sum_{\boldsymbol{\alpha}'\in\mathbb{N}_0^d}\left\langle \psi_{\boldsymbol{\alpha}'}, \frac{f_{+\alpha j} - f_{-\alpha j}}{2}\right\rangle^2 .
\end{aligned}$$

Since the functions $f_{-\alpha j}$ and $f_{+\alpha j}$ are constant in $x_j$ on $I_{\alpha\mathbf{e}_j}$ and vanish outside, we only need to consider summands with coarser resolution $\lambda_j' < \lambda$ in that coordinate, and where the support of $\psi_{\boldsymbol{\alpha}}$ contains the support of $f_{\pm\alpha j}$. That is the case for $\kappa_j' = \lfloor 2^{\lambda_j'-\lambda}\kappa\rfloor$ with $\lambda_j' = -\infty, 0, \ldots, \lambda-1$. For these indices we have

$$\left\langle \psi_{\boldsymbol{\alpha}'}, \frac{f_{+\alpha j} - f_{-\alpha j}}{2}\right\rangle^2 = 2^{\max\{0,\lambda_j'\}-\lambda}\,\langle \psi_{\boldsymbol{\alpha}''}, f\rangle^2 = 2^{\max\{0,\lambda_j'\}-\lambda}\,\tilde{f}^2(\boldsymbol{\alpha}'')\,,$$

where $\alpha_{j'}'' = \alpha_{j'}'$ for $j' \neq j$, and $\alpha_j'' = \alpha$. Hence we obtain

$$\tilde{f}(\alpha\,\mathbf{e}_j) \geq 2^{\lambda/2}\underbrace{\left(2^{-\lambda} + \sum_{l=0}^{\lambda-1} 2^{l-\lambda}\right)}_{=1}\sum_{\substack{\boldsymbol{\alpha}''\in\mathbb{N}_0^d \\ \alpha_j''=\alpha}} \tilde{f}^2(\boldsymbol{\alpha}'')\,. \tag{4.4.3}$$

Based on this relation between the wavelet coefficients, we can estimate

$$\begin{aligned}
1 = \|f\|_{L_2}^2 &= \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^d} \tilde{f}^2(\boldsymbol{\alpha}) \\
&\geq \sum_{j=1}^{d}\sum_{\lambda=0}^{r-1}\sum_{\kappa=0}^{2^\lambda-1} \tilde{f}^2((2^\lambda+\kappa)\,\mathbf{e}_j) \\
&\geq \sum_{j=1}^{d}\sum_{\lambda=0}^{r-1}\left(2^{-\lambda/2}\sum_{\kappa=0}^{2^\lambda-1}\tilde{f}((2^\lambda+\kappa)\,\mathbf{e}_j)\right)^2 \\
&\overset{(4.4.3)}{\geq} \sum_{j=1}^{d}\sum_{\lambda=0}^{r-1}\left(\sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d \\ \lambda_j=\lambda}}\tilde{f}^2(\boldsymbol{\alpha})\right)^2 .
\end{aligned}$$

Taking the square root, and using the norm estimate $\|\mathbf{x}\|_1 \leq \sqrt{m}\,\|\mathbf{x}\|_2$ for $\mathbf{x}\in\mathbb{R}^m$,

here with $m = d\,r$, we get

$$\begin{aligned} 1 &\geq \frac{1}{\sqrt{d\,r}} \sum_{j=1}^{d} \sum_{\lambda=0}^{r-1} \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d \\ \lambda_j=\lambda}} \tilde{f}^2(\boldsymbol{\alpha}) \\ &\geq \frac{1}{\sqrt{d\,r}} \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d \\ \boldsymbol{\lambda}<r}} |\boldsymbol{\alpha}|_0\, \tilde{f}^2(\boldsymbol{\alpha}) \\ &\geq \frac{k+1}{\sqrt{d\,r}} \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d \\ |\boldsymbol{\alpha}|_0>k \\ \boldsymbol{\lambda}<r}} \tilde{f}^2(\boldsymbol{\alpha})\,. \end{aligned}$$

This proves the lemma. □

**Theorem 4.22.** *For the algorithm $A_{n,k,r} = (A^{\omega}_{n,k,r})_{\omega\in\Omega}$ we have the error bound*

$$e(A_{n,k,r}, F^d_{\mathrm{mon}\pm} \hookrightarrow L_1([0,1]^d)) \leq \frac{d}{2^{r+1}} + \sqrt{\frac{\sqrt{d\,r}}{k+1} + \frac{\exp[k(1+\log\frac{d}{k} + (\log 2)\,r)]}{n}}\,.$$

*Given $0 < \varepsilon < 1$, the $\varepsilon$-complexity for the Monte Carlo approximation of monotone functions is bounded by*

$$n^{\mathrm{ran}}(\varepsilon, F^d_{\mathrm{mon}\pm} \hookrightarrow L_1([0,1]^d), \Lambda^{\mathrm{std}}) \;\leq\; \min\left\{ \exp\left[C\,\frac{\sqrt{d}}{\varepsilon^2}\,\left(1+\left(\log\frac{d}{\varepsilon}\right)^{3/2}\right)\right], \exp\left[d\,\log\frac{d}{2\varepsilon}\right]\right\},$$

*with some numerical constant $C > 0$. In particular, the curse of dimensionality does* not *hold for the randomized $L_1$-approximation of monotone functions.*

*Proof.* Since we only take certain wavelet coefficients until a resolution $r$ into account, the output will be a function that is constant on each of $2^{rd}$ subcubes $C_{r\mathbf{1},\boldsymbol{\kappa}}$ where $\boldsymbol{\kappa} \in \{0,\dots,2^r-1\}^d$. The algorithm actually approximates

$$f_r := \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d \\ \boldsymbol{\lambda}<r}} \tilde{f}(\boldsymbol{\alpha})\,\psi_{\boldsymbol{\alpha}}\,.$$

Since on the one hand, the Haar wavelets are constant on each of the $2^{rd}$ subcubes, and on the other hand, we have $2^{rd}$ wavelets up to this resolution, the function $f_r$ averages the function $f$ on each of the subcubes. That is, for $\mathbf{X}, \mathbf{X}' \sim \mathrm{unif}\, C_{r\mathbf{1},\boldsymbol{\kappa}}$ we have

$$\mathbb{E}\,|f(\mathbf{X}) - f_r(\mathbf{X})| = \mathbb{E}\,|f(\mathbf{X}) - \mathbb{E}'\,f(\mathbf{X}')| \leq \frac{1}{2}\left[\sup_{\mathbf{x}\in C_{r\mathbf{1},\boldsymbol{\kappa}}} f(\mathbf{x}) - \inf_{\mathbf{x}\in C_{r\mathbf{1},\boldsymbol{\kappa}}} f(\mathbf{x})\right]. \quad (4.4.4)$$

Following the same arguments as in the proof of Theorem 4.2, we group the subcubes into diagonals, each diagonal being uniquely represented by a $\boldsymbol{\kappa}$ with at least one 0-entry. By monotonicity, summing up (4.4.4) for all cubes of a diagonal, we obtain the upper bound $\frac{1}{2}$. Now that there are $2^{rd} - 2^{r(d-1)} \leq d\, 2^{r(d-1)}$ diagonals, and the volume of each subcube is $2^{-rd}$, we obtain the estimate

$$\|f - f_r\|_{L_1} \leq \frac{d}{2^{r+1}}\,. \tag{4.4.5}$$

Surprisingly, the fact that the wavelet basis functions have a small support, actually helps to keep the error for estimating the wavelet coefficients small. Exploiting independence and unbiasedness (compare (4.4.2)), for $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ we have

$$\begin{aligned}
\mathbb{E}[\tilde{f}(\boldsymbol{\alpha}) - \tilde{g}(\boldsymbol{\alpha})]^2 &= \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}[\tilde{f}(\boldsymbol{\alpha}) - \psi_{\boldsymbol{\alpha}}(\mathbf{X}_i)\, f(\mathbf{X}_i)]^2 \\
&= \frac{1}{n} \Big( \mathbb{E}[\psi_{\boldsymbol{\alpha}}(\mathbf{X}_i)\, f(\mathbf{X}_i)]^2 - \tilde{f}^2(\boldsymbol{\alpha}) \Big) \\
&\leq \frac{1}{n} \underbrace{\mathbb{P}\{\mathbf{X}_1 \in C_{\boldsymbol{\alpha}}\}}_{=2^{-\lambda}} \mathbb{E}\big[ \underbrace{(\psi_{\boldsymbol{\alpha}}(\mathbf{X}_1)\, f(\mathbf{X}_1))^2}_{\in[0,2^{\lambda}]} \mid \mathbf{X}_1 \in C_{\boldsymbol{\alpha}} \big] \\
&\leq \frac{1}{n}\,.
\end{aligned} \tag{4.4.6}$$

Then by (4.4.5), Lemma 4.21, and (4.4.6), the expected distance between input $f$ and output $g$ is

$$\begin{aligned}
\mathbb{E}\,\|f - g\|_{L_1} &\leq \|f - f_r\|_{L_1} + \mathbb{E}\,\|f_r - g\|_{L_2} \\
&\leq \|f - f_r\|_{L_1} + \Bigg( \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d \\ |\boldsymbol{\alpha}|_0 > k \\ \boldsymbol{\lambda} < r}} \tilde{f}(\boldsymbol{\alpha})^2 + \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d \\ |\boldsymbol{\alpha}|_0 \leq k \\ \boldsymbol{\lambda} < r}} \mathbb{E}[\tilde{f}(\boldsymbol{\alpha}) - \tilde{g}(\boldsymbol{\alpha})]^2 \Bigg)^{\frac{1}{2}} \\
&\leq \frac{d}{2^{r+1}} + \sqrt{\frac{\sqrt{d\,r}}{k+1} + \frac{\#A}{n}}\,,
\end{aligned}$$

where

$$A := \{\boldsymbol{\alpha} \in \mathbb{N}_0^d \mid |\boldsymbol{\alpha}|_0 \leq k \text{ and } \boldsymbol{\lambda} < r\}\,.$$

Using Lemma B.1, we can estimate the size of $A$ for $k \in \{1, \ldots, d\}$,

$$\#A = \sum_{l=0}^{k} \binom{d}{l} (2^r - 1)^l \leq 2^{r\,k} \left( \frac{\mathrm{e}\,d}{k} \right)^k\,.$$

This gives us the error bound for the Monte Carlo method $A_{n,k,r}$.

Choosing the resolution $r := \lceil \log_2 \frac{d}{\varepsilon} \rceil$ will bound the first term $2^{-(r+1)}\, d \leq \varepsilon/2$. Taking $k := \big\lfloor 8 \sqrt{d\,(1 + \log_2 \frac{d}{\varepsilon})}/\varepsilon^2 \big\rfloor \overset{!}{\leq} d$ then guarantees $\sqrt{d\,r}/(k+1) \leq \varepsilon^2/8$. Finally, $\#A/n$ can be bounded from above by $\varepsilon^2/8$ if we choose

$$n := \left\lceil \frac{8}{\varepsilon^2} \exp\left( \frac{8 \sqrt{d\,(1 + \log_2 \frac{d}{\varepsilon})}}{\varepsilon^2} \left( 1 + \log \frac{d}{k} + \log \frac{2\,d}{\varepsilon} \right) \right) \right\rceil\,.$$

By this choice we obtain the error bound $\varepsilon$ we aimed for.

Note that if $\varepsilon$ is too small, we can only choose $k = d$ for the algorithm $A_{n,k,r}$. In this case, for the approximation of $f_r$, we would take $2^{rd}$ wavelet coefficients into account, $n$ would become much bigger in order to achieve the accuracy we aim for. Instead, one can approximate $f$ directly via the deterministic algorithm $A_m^d$ from Theorem 4.2, which is based on $m^d$ function values on a regular grid. The worst case error is bounded by $e(A_m^d) \le d/(2(m+1))$. Taking $m := 2^r - 1$, this gives the same bound that we have for the accuracy at which $f_r$ approximates $f$, see (4.4.5). So for small $\varepsilon$ we take the deterministic upper bound

$$n^{\mathrm{det}}(\varepsilon, \mathrm{APP}, F_{\mathrm{mon}}^d, \Lambda^{\mathrm{std}}) \le \exp\left(d \, \log \frac{d}{2\,\varepsilon}\right),$$

compare Remark 4.3. □

**Remark 4.23** (Non-linear algorithm with improved $\varepsilon$-dependency)**.** It is rather unpleasant that the estimate in Theorem 4.22 depends exponentially on $\varepsilon^{-2}$, at least for $\varepsilon_d \succeq 1/\sqrt[4]{d}$. In other words, for $d$-dependent error tolerances $\varepsilon_d \asymp 1/\sqrt[4]{d}$, the cardinality $n = n(\varepsilon_d, d)$ of $A_{n,k,r}$ (with appropriately chosen parameters) depends exponentially on $d$. However, there is a way to improve the $\varepsilon$-dependency at the price of losing the linearity of the algorithm.

For the subclass of sign-valued monotone functions

$$F_{\mathrm{mon}\{\pm\}}^d := \{f : [0,1]^d \to \{-1,+1\} \mid f \in F_{\mathrm{mon}\pm}^d\},$$

we can modify the algorithm in a way similar to the Boolean setting, the new version $\tilde{A}_{n,k,r}$ now returning an output $\tilde{g} := \operatorname{sgn} g$. For $f \in F_{\mathrm{mon}\{\pm\}}^d$ we can estimate

$$\begin{aligned}\mathbb{E}\,\|f - \tilde{g}\|_{L1} &\le \|f - f_r\|_{L_1} + 2\,\|f_r - g\|_{L_2}^2 \\ &\le \frac{d}{2^{r+1}} + 2\,\frac{\sqrt{dr}}{k+1} + 2\,\frac{\#A}{n},\end{aligned}$$

and for the restricted input set $F_{\mathrm{mon}\{\pm\}}^d \subset F_{\mathrm{mon}\pm}^d$ we obtain the complexity bound

$$n^{\mathrm{ran}}(\varepsilon, F_{\mathrm{mon}\{\pm\}}^d \hookrightarrow L_1[0,1]^d, \Lambda^{\mathrm{std}}) \le \exp\left[C'\,\frac{\sqrt{d}}{\varepsilon}\left(1 + \left(\log\frac{d}{\varepsilon}\right)^{3/2}\right)\right], \qquad (4.4.7)$$

with a numerical constant $C' > 0$.

This complexity bound holds actually for the whole class $F_{\mathrm{mon}\pm}^d$. Indeed, any bounded monotone function $f \in F_{\mathrm{mon}\pm}^d$ can be written as an integral composition of sign-valued functions $f_t := \operatorname{sgn}(f - t) \in F_{\mathrm{mon}\{\pm\}}^d$,

$$f = \frac{1}{2}\int_{-1}^{1} f_t \,\mathrm{d}t\,.$$

We define a new algorithm $\bar{A}_{n,k,r}$ by

$$\bar{A}_{n,k,r}^{\omega}(f) := \frac{1}{2}\int_{-1}^{1} \tilde{A}_{n,k,r}^{\omega}(f_t)\,\mathrm{d}t\,.$$

Note that the information needed for the computation of $\tilde{A}^{\omega}_{n,k,r}(f_t)$ can be derived from the same information mapping applied to $f$ directly since the algorithm is non-adaptive. We can write $\bar{A}^{\omega}_{n,k,r} = \bar{\phi}^{\omega}_{n,k,r} \circ N^{\omega}$ in contrast to $A^{\omega}_{n,k,r} = \phi^{\omega}_{n,k,r} \circ N^{\omega}$. By the triangle inequality we get

$$\begin{aligned} e(\bar{A}_{n,k,r}, f) &= \mathbb{E}\,\|f - \bar{A}^{\omega}_{n,k,r}(f)\| \\ &\le \frac{1}{2}\int_{-1}^{1} \mathbb{E}\,\|f_t - \tilde{A}^{\omega}_{n,k,r}(f_t)\|\,\mathrm{d}t \\ &\le e(\tilde{A}_{n,k,r}, F^d_{\mathrm{mon}\{\pm\}})\,. \end{aligned}$$

**Remark 4.24** (Realization of the non-linear method $\bar{A}_{n,k,r}$)**.** The resulting algorithm $\bar{A}_{n,k,r}$ is not linear anymore. It is an interesting question how much the *combinatory cost* for $\bar{\phi}_{n,k,r}$ differs from the cost for $\phi_{n,k,r}$. For the model of computation we refer to the book on IBC of Traub et al. [73, p. 30], and to Novak and Woźniakowski [55, Sec 4.1.2].

The cost for computing the **linear representation** $\phi_{n,k,r}$ is dominated by the following operations:

- Each sample $f(\mathbf{X}_i)$ contributes to $\sum_{l=0}^{k}\binom{d}{l}(r+1)^l \le \left(\frac{\mathrm{e}\,(r+1)\,d}{k}\right)^k$ wavelet coefficients. The relevant indices $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ can be determined effectively based on the binary representation of $\mathbf{X}_i$.
- Compute the linear combination of $\#A = \sum_{l=0}^{k}\binom{d}{l}(2^r-1)^l \le \left(\frac{\mathrm{e}\,2^r\,d}{k}\right)^k$ wavelets.

The first part is the most costly part with more than $n$ operations needed. If the parameters are chosen according to Theorem 4.22, the second part only needs about $n\,\varepsilon^2/4$ operations in $L_2([0,1]^d)$. We can roughly summarize the cost as $n \preceq \operatorname{cost}\phi_{n,k,r} \preceq n^2$.

For the **non-linear representation** $\bar{\phi}_{n,k,r}$, proceed as follows:

- Sort the information $(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_n, y_n)$ such that $y_1 \le y_2 \le \ldots \le y_n$.
- Define $y_0 := -1$ and $y_{n+1} := +1$, and use the representation

$$\bar{\phi}^{\omega}_{n,k,r}(\mathbf{y}) = \frac{1}{2}\sum_{i=0}^{n}(y_{i+1}-y_i)\,\operatorname{sgn}\underbrace{\phi^{\omega}_{n,k,r}(\overbrace{-1,\ldots,-1}^{i\text{ times}},\ \overbrace{1,\ldots,1}^{(n-i)\text{ times}})}_{=:g_i}\,.$$

Here, the cost for computing $g_0$ is the cost for computing $\phi_{n,k,r}$. For $i = 1, \ldots, n$, we obtain $g_i$ from modifying $g_{i-1}$, indeed, by linearity of $\phi_{n,k,r}$ we have

$$g_i := g_{i-1} - 2\,\phi^{\omega}_{n,k,r}(\mathbf{e}_i)\,.$$

Since for $\phi^{\omega}_{n,k,r}(\mathbf{e}_i)$ we only need to take $\sum_{l=0}^{k}\binom{d}{l}(r+1)^l \le \left(\frac{\mathrm{e}\,(r+1)\,d}{k}\right)^k$ wavelet coefficients into account, doing this $n$ times, the cost for computing $g_0, \ldots, g_n$ is only twice the cost of $\phi_{n,k,r}$. (Here, however, we need more operations in $L_2([0,1]^d)$, and

less operations in $\mathbb{R}$, but we assumed them to have the same cost, no matter how realistic that is.) The signum operator and the final sum contribute to the cost only linearly in $n$. Ordering the information has an expected cost of $n \log n$, which is likely to be dominated by the cost of $\phi_{n,k,r}$. Assuming this, we obtain

$$\operatorname{cost} \phi_{n,k,r} \asymp \operatorname{cost} \bar{\phi}_{n,k,r} \,.$$

Heinrich and Milla [26, Sec 6.2] pointed out that for problems with functions as output, the interesting question is not about a complete picture of the output $\phi(\mathbf{y})$, but about effective computation of approximate function values $[\phi(\mathbf{y})](\mathbf{x})$ on demand. In our situation it makes sense to distinguish between pre-processing operations and operations on demand.

For the **linear representation** $\phi_{n,k,r}$ we have

*Pre-processing:* Compute and store the wavelet coefficients that are needed for the output.

*On demand:* Compute $[\phi^{\omega}_{n,k,r}(\mathbf{y})](\mathbf{x})$.
(Only $\sum_{l=0}^{k} \binom{d}{l}(r+1)^l$ wavelet coefficients are relevant to $[\phi^{\omega}_{n,k,r}(\mathbf{y})](\mathbf{x})$.)

The pre-processing is approximately as expensive as the cost with the above computational model, computation on demand is significantly cheaper.

For the **non-linear representation** $\bar{\phi}_{n,k,r}$ we have

*Pre-processing:* Rearrange the information.
Store the wavelet coefficients needed for $g_0$.

*On demand:* Compute $g_0(\mathbf{x})$.
(For this, $\sum_{l=0}^{k} \binom{d}{l}(r+1)^l$ wavelet coefficients are relevant.)
In order to compute $[\bar{\phi}^{\omega}_{n,k,r}(\mathbf{y})](\mathbf{x})$, we need in particular the values $[\phi^{\omega}_{n,k,r}(\mathbf{e}_i)](\mathbf{x})$ for $i = 1, \dots, n$. These can be determined in a very effective way.[21]

[21] Observe that

$$n\,[\phi^{\omega}_{n,k,r}(\mathbf{e}_i)](\mathbf{x}) = \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d \\ |\boldsymbol{\alpha}|_0\le k \\ |\boldsymbol{\alpha}|_\infty<2^r}} \psi_{\boldsymbol{\alpha}}(\mathbf{X}_i)\,\psi_{\boldsymbol{\alpha}}(\mathbf{x}) = \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d \\ |\boldsymbol{\alpha}|_0\le k \\ |\boldsymbol{\alpha}|_\infty<2^r}} \prod_{j=1}^{d} h_{\alpha_j}(\mathbf{X}_i(j))\,h_{\alpha_j}(x_j) = \sum_{\substack{\boldsymbol{\beta}\in\{0,1\}^d \\ |\beta|_1\le k}} \mathbf{Z}^{\boldsymbol{\beta}}\,,$$

where $Z_j := \sum_{\alpha_j=1}^{2^r-1} h_{\alpha_j}(\mathbf{X}_i(j))\,h_{\boldsymbol{\alpha}_j}(x_j)$. It is readily checked that

$$Z_j = \begin{cases} 2^r - 1 & \text{if } \lfloor 2^r\,\mathbf{X}_i(j)\rfloor = \lfloor 2^r\,x_j\rfloor, \\ -1 & \text{else,} \end{cases}$$

so a comparison of the first $r$ digits of the binary representation of $\mathbf{X}_i(j)$ and $x_j$ is actually enough for determining $Z_j$. In the end, we only need the number $b$ of coordinates where $\lfloor 2^r\,\mathbf{X}_i(j)\rfloor = \lfloor 2^r\,x_j\rfloor$, and obtain

$$n\,[\phi^{\omega}_{n,k,r}(\mathbf{e}_i)](\mathbf{x}) = \sum_{l=0}^{b\wedge k} \binom{b}{l} (2^r-1)^l \sum_{m=0}^{(d-b)\wedge(k-l)} \binom{d-b}{m} (-1)^m =: \chi(b) \in \mathbb{Z}\,.$$

These values $\chi(b)$ are needed for $b \in \{0, \dots, d\}$. Since they only depend on parameters of the algorithm, they can be prepared before any information was collected.

For the pre-processing there is no big difference from the setting before. The part "on demand" it is a little cheaper than the pre-processing part, however, we need more than $n$ operations. Hence a linear algorithm with the same information cardinality on the one hand is less costly, on the other hand the error is larger.

We conclude that it depends on $\varepsilon$ and the ratio of information cost versus combinatory cost whether the linear or the non-linear algorithm should be preferred.

**Remark 4.25** (Deterministic methods for $\Lambda^{\mathrm{all}}$)**.** If we are allowed asking the oracle for wavelet coefficients directly, the same algorithmic idea is implementable as a deterministic method and less information is needed. (This reduction of the complexity is by a factor $\varepsilon^2/8$ or $\varepsilon/4$.) In particular, the curse of dimensionality does *not* hold in the deterministic setting with $\Lambda^{\mathrm{all}}$.

It is an open problem whether similar lower bounds to those from Section 4.3 can be found for $\Lambda^{\mathrm{all}}$. The proof technique of Theorem 4.8, however, will not work for $\Lambda^{\mathrm{all}}$, because one could choose a functional that injectively maps all possible Boolean functions onto the real line, and thus identify any given function by just one measurement.

Indeed, one measurement is sufficient. Let $\mathbf{x} \in \{0,1\}^d$ be the binary representation of an integer,

$$b(\mathbf{x}) := \sum_{j=1}^{d} x_j \, 2^{j-1} \in \mathbb{N}_0.$$

Define the functional $L_1$ for Boolean functions $f : \{0,1\}^d \to \{0,1\}$ by

$$L_1(f) := \sum_{\mathbf{x} \in \{0,1\}^d} f(\mathbf{x}) 2^{b(\mathbf{x})} \, .$$

This functional maps $G^d_{\circ\!|}$ bijectively onto the set $\{0, \ldots, 2^{2^d} - 1\}$ of $2^d$-bit representable integers. This shows that $\Lambda^{\mathrm{all}}$ is an inappropriate model for the approximation of Boolean functions.

# Appendix A

# On Gaussian Measures

Let $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$ denote the cumulative distribution function of a standard Gaussian variable.
Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a standard Gaussian vector in $\mathbb{R}^m = \ell_2^m$, i.e. the $X_i$ are iid standard Gaussian variables. For any linear mapping $J : \ell_2^m \to \widetilde{F}$ with $\widetilde{F}$ being a Banach space, then $J\mathbf{X}$ is a zero-mean Gaussian vector in $\widetilde{F}$. For the norm of a vector $\mathbf{x} \in \ell_2^m$ we write $\|\mathbf{x}\|_2$. The operator norm of $J$ shall be denoted $\|J\|_{2\to F}$.

## A.1 Comparison of Gaussian Measures

**Lemma A.1.** *Let* $\mathbf{Y}$ *and* $\mathbf{Z}$ *be independent zero-mean Gaussian vectors in a normed space* $\widetilde{F}$*. Then*

$$\mathbb{E}\,\|\mathbf{Y}\|_F \leq \mathbb{E}\,\|\mathbf{Y} + \mathbf{Z}\|_F\,.$$

*Proof.* Due to symmetry of zero-mean Gaussian measures, the distributions of $\mathbf{Y} + \mathbf{Z}$ and $\mathbf{Y} - \mathbf{Z}$ are identical. Hence, by the triangle inequality,

$$\begin{aligned}\mathbb{E}\,\|\mathbf{Y}\|_F &= \mathbb{E}\,\left\|\tfrac{1}{2}\textstyle\sum_{\sigma=\pm1}(\mathbf{Y} + \sigma\mathbf{Z})\right\|_F \\ &\leq \tfrac{1}{2}\textstyle\sum_{\sigma=\pm1}\mathbb{E}\,\|\mathbf{Y} + \sigma\mathbf{Z}\|_F \\ &= \mathbb{E}\,\|\mathbf{Y} + \mathbf{Z}\|_F\,.\end{aligned}$$

□

**Lemma A.2.** *Let* $X_1, \ldots, X_m$ *be independent standard Gaussian random variables,* $f_1, \ldots, f_m$ *be elements in a normed vector space* $\widetilde{F}$*, and* $a'_k \geq a_k \geq 0$ *be real numbers. Then*

$$\mathbb{E}\left\|\sum_{k=1}^{m} a_k\, X_k\, f_k\right\|_F \leq \mathbb{E}\left\|\sum_{k=1}^{m} a'_k\, X_k\, f_k\right\|_F\,.$$

*Proof.* Take another $m$ independent standard Gaussian random variables $X'_1, \ldots, X'_m$. Then the $a_k\, X_k + \sqrt{a_k'^2 - a_k^2}\, X'_k$ are identically distributed to $a'_k\, X_k$. Thus, applying

Lemma A.1, we obtain

$$\mathbb{E}\left\|\sum_{k=1}^{m} a_k\, X_k\, f_k\right\|_F \le \mathbb{E}\left\|\sum_{k=1}^{m} a_k\, X_k\, f_k + \sum_{k=1}^{m}\sqrt{a_k'^2 - a_k^2}\, X_k' f_k\right\|_F = \mathbb{E}\left\|\sum_{k=1}^{m} a_k'\, X_k\, f_k\right\|_F .$$

□

There are several other known comparison principles for Gaussian vectors, mainly based on covariance comparisons, and concerning the expected maximum component instead of arbitrary norms, see for example Lifshits [43, pp. 186–192]. For Gaussian fields one important result is Slepian's inequality, see e.g. Adler [1, Cor 2.4].

## A.2 Restrictions of Gaussian Measures

With $P$ being an orthogonal projection in $\ell_2^m$, the random vector $JP\mathbf{X}$ can be interpreted as a restriction of the Gaussian measure to the subspace $\operatorname{img} JP \subseteq \widetilde{F}$. We start with a rather simple comparison of the expected norms.

**Lemma A.3.** *For orthogonal projections $P$ on $\ell_2^m$ we have*

$$\mathbb{E}\,\|JP\mathbf{X}\|_F \le \mathbb{E}\,\|J\mathbf{X}\|_F .$$

*Proof.* Due to orthogonality, $\mathbf{Y} := JP\mathbf{X}$ and $\mathbf{Z} := J(\mathrm{id} - P)\mathbf{X}$ are stochastically independent. The claim follows from Lemma A.1. □

As an application we have a bound for the operator norm of $J$.

**Lemma A.4.** *For any linear operator $J : \ell_2^m \to \widetilde{F}$ we have*

$$\mathbb{E}\,\|J\mathbf{X}\|_F \ge \sqrt{\tfrac{2}{\pi}}\,\|J\|_{2\to F} .$$

*Proof.* Applying Lemma A.3 to rank-1 orthogonal projections, we obtain

$$\mathbb{E}\,\|J\mathbf{X}\|_F \ge \sup_{\substack{P \text{ orth. Proj.}\\ \operatorname{rk} P=1}} \mathbb{E}\,\|JP\mathbf{X}\|_F = \|J\|_{2\to F}\,\mathbb{E}\,|X_1| = \sqrt{\tfrac{2}{\pi}}\,\|J\|_{2\to F} .$$

□

## A.3 Deviation Estimates for Gaussian Measures

Corollary A.6 is a deviation result for norms of Gaussian vectors known from Pisier [65, Thm 2.1, 2.2, Rem p. 180/181]. It has been used in this form by Heinrich [22, Prop 1] for his proof on lower bounds for the Monte Carlo error via Gaussian measures, which is reproduced in Chapter 2. The deviation result for the norms of Gaussian vectors is a consequence of the following proposition.

**Proposition A.5.** *Let $f : \ell_2^m \to \mathbb{R}$ be a Lipschitz function with Lipschitz constant $L$. Then for $t > 0$ we have*

$$\mathbb{P}\{f(\mathbf{X}) - \mathbb{E}\, f(\mathbf{X}) > t\} \leq \exp\left(-\frac{t^2}{2L^2}\right).$$

For this result several proofs are known. One way uses stochastic integration, a rather short description of this approach can be found in Pisier [65, Rem p. 180/181]. A more direct proof is contained in Adler and Taylor [2, Lem 2.1.6]. Both methods require higher smoothness for $f$ first, an assumption that can be removed in the end with Fatou's inequality. For a simpler proof with a slightly worse constant in the exponent, see Pisier [65, p. 176–178].

**Corollary A.6.** *For $\lambda > 1$ and with $\rho := \mathbb{E}\,\|J\mathbf{X}\|_F / \|J\|_{2\to F}$ we have*

$$\mathbb{P}\{\|J\mathbf{X}\|_F > \lambda\, \mathbb{E}\,\|J\mathbf{X}\|_F\} \leq \exp\left(-\frac{(\lambda-1)^2\,\rho^2}{2}\right) \leq \exp\left(-\frac{(\lambda-1)^2}{\pi}\right).$$

*Proof.* We define $f : \ell_2^m \to \mathbb{R}$ by

$$f(\mathbf{x}) := \|J\mathbf{x}\|_F\,.$$

This function is Lipschitz continuous with Lipschitz constant

$$L = \sup_{\substack{\mathbf{x},\mathbf{z}\in\ell_2^m \\ \mathbf{x}\neq\mathbf{z}}} \frac{|f(\mathbf{x}) - f(\mathbf{z})|}{\|\mathbf{x}-\mathbf{z}\|_2} \overset{\Delta\text{-ineq.}}{\leq} \sup_{\substack{\mathbf{x},\mathbf{z}\in\ell_2^m \\ \mathbf{x}\neq\mathbf{z}}} \frac{\|J(\mathbf{x}-\mathbf{y})\|_F}{\|\mathbf{x}-\mathbf{z}\|_2} = \|J\|_{2\to F}\,,$$

so that we can apply Proposition A.5 with $t = (\lambda-1)\,\mathbb{E}\,\|J\mathbf{X}\|_F$. Actually, we have equality $L = \|J\|_{2\to F}$, to see this, consider the LHS with the supremum for $\mathbf{x} \neq \mathbf{0} = \mathbf{z}$.

For this kind of estimate it is sufficient to bound $\|J\|_{2\to F}^2$ from above, or the ratio $\rho$ from below. In the most general way, $\rho \geq \sqrt{2/\pi}$, see Lemma A.4. □

## A.4 Gaussian Vectors in Sequence Spaces

The next two lemmas are needed for the proof of Lemma A.9 which follows the lines of Pisier [66, Lem 4.14]. In there the inequalities of Lemma A.7 and A.8 have been mentioned without giving a proof or detailed information about the constants.

**Lemma A.7** (Tail estimate)**.** *There exists a constant $\alpha > 0$ such that*

$$\mathbb{P}\left\{|X| > \alpha\sqrt{\log x}\right\} \geq \frac{1}{x} \quad \text{for } x \geq \mathrm{e}, \tag{A.4.1}$$

*where $X$ is a standard Gaussian random variable.*
*The value $\alpha = \Phi^{-1}\left(1 - \frac{1}{2\mathrm{e}}\right) = -\Phi^{-1}\left(\frac{1}{2\mathrm{e}}\right) = 0.90045...$ is optimal, providing equality for $x = \mathrm{e}$. Here, $\Phi$ denotes the cumulative distribution function of the standard normal distribution.*

*Proof.* Having chosen $\alpha$ as above, we consider the left-hand side of the inequality to be shown,

$$\begin{aligned}\mathrm{LHS} = \mathbb{P}\left\{|X| > \alpha\sqrt{\log x}\right\} &= \sqrt{\frac{2}{\pi}}\int_{\alpha\sqrt{\log x}}^{\infty} \exp\left(-\frac{t^2}{2}\right)\,\mathrm{d}t \\ [t = \alpha\sqrt{\log s}] \quad &= \frac{\alpha}{\sqrt{2\pi}}\int_x^{\infty} \frac{s^{-(\alpha^2/2+1)}}{\sqrt{\log s}}\,\mathrm{d}s\end{aligned}$$

Its derivative is

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{LHS} = -\frac{\alpha}{2\pi}\,\frac{x^{-(\alpha^2/2+1)}}{\sqrt{\log x}}\,,$$

whereas for the right-hand side we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{RHS} = -\frac{1}{x^2}\,.$$

The derivatives of both sides agree iff

$$\frac{\alpha^2}{2\pi}x^{2-\alpha^2} = \log x\,. \tag{A.4.2}$$

Since $\alpha < 1$, we have that $x^{2-\alpha^2}$ is convex, whereas $\log x$ is concave for $x > 0$, so there are no more than two points $0 < x_1 < x_2$ that solve (A.4.2). Comparing both sides of (A.4.2), we obtain:

$$\begin{aligned}\frac{\alpha}{\sqrt{2\pi}} > 0 \quad &\text{for } x = 1, \\ \frac{\alpha^2}{2\pi}\mathrm{e}^{2-\alpha^2} \overset{\alpha<1}{<} \frac{\mathrm{e}}{2\pi} < 1 \quad &\text{for } x = \mathrm{e}, \text{ and} \\ \frac{\alpha^2}{2\pi}x^{2-\alpha^2} > \log x \quad &\text{for sufficiently large } x, \text{ say } x > x_2.\end{aligned}$$

Therefore $1 < x_1 < \mathrm{e} < x_2$.

Now for (A.4.1), by choice of $\alpha$, we have

$$\begin{aligned}\mathrm{LHS} = \mathrm{RHS} \qquad &\text{for } x = \mathrm{e}, \\ \lim_{x\to\infty}\mathrm{LHS} = \lim_{x\to\infty}\mathrm{RHS} &= 0\,.\end{aligned}$$

With

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{LHS}\,|_{x=\mathrm{e}} &= -\frac{\alpha}{\sqrt{2\pi}}\,\mathrm{e}^{-(\alpha^2/2+1)} \approx -0.088107 \\ &> \frac{\mathrm{d}}{\mathrm{d}x}\mathrm{RHS}\,|_{x=\mathrm{e}} = -\mathrm{e}^{-2} \approx -0.135335\,,\end{aligned}$$

and only having one point $x = x_2 > \mathrm{e}$ where the difference LHS − RHS is locally extreme, we obtain that LHS ≥ RHS for $\mathrm{e} \le x < \infty$. □

**Lemma A.8.** *There exists a constant $K > 0$ such that for all $1 \le q < \infty$ we have*

$$(\mathbb{E}\,|X|^q)^{1/q} \le K\sqrt{q}\,,$$

*where $X$ is a standard Gaussian variable.*

*Proof.* We can write

$$\gamma(q) := (\mathbb{E}\,|X|^q)^{1/q} = \left(\sqrt{\frac{2}{\pi}}\int_0^\infty x^q \exp\left(-\frac{x^2}{2}\right)\,\mathrm{d}x\right)^{1/q}.$$

Given $\alpha \in (0,1)$, one can find $c > 0$ such that

$$x^q \exp\left(-\frac{x^2}{2}\right) \le c\,x \exp\left(-\alpha\,\frac{x^2}{2}\right),$$

in detail, $c = \left(\dfrac{q-1}{\mathrm{e}\,(1-\alpha)}\right)^{(q-1)/2}$ is optimal with equality holding in $x = \sqrt{\dfrac{q-1}{1-\alpha}}$. We obtain

$$\int_0^\infty x^q \exp\left(-\frac{x^2}{2}\right)\,\mathrm{d}x \le c\int_0^\infty x \exp\left(-\alpha\,\frac{x^2}{2}\right)\,\mathrm{d}x = \frac{c}{\alpha} = \frac{1}{\alpha}\left(\frac{q-1}{\mathrm{e}\,(1-\alpha)}\right)^{(q-1)/2}.$$

This estimate can be minimized by choosing $\alpha = \dfrac{2}{q+1}$, so

$$\gamma(q) \le \underbrace{\frac{1}{\sqrt{\mathrm{e}}}\left(\frac{\mathrm{e}\,(q+1)}{2\pi}\right)^{1/(2q)}\sqrt{1+\frac{1}{q}}}_{=:K(q)}\sqrt{q}\,.$$

We can carry out a final estimate (finding the extreme point by $0 \overset{!}{=} \frac{\mathrm{d}}{\mathrm{d}q}\log K(q)$),

$$K := \sup_{q\ge 1} K(q) = K\left(\frac{2\pi}{\mathrm{e}} - 1\right) = \sqrt{\frac{2\pi}{\mathrm{e}\,(2\pi - \mathrm{e})}} \approx 0.805228\,.$$

This value for $K$ is not optimal because for $q > 1$ our estimate is rough. But it is close to optimal since $\gamma(1) = \sqrt{2/\pi} \approx 0.797885$, i.e. if one could show that the lemma is true with $K = \sqrt{2/\pi}$, one would have a sharp estimate with equality holding for $q = 1$. Numerical calculations strongly encourage this conjecture. □

Now we can state and prove the norm estimates for Gaussian vectors, the proof (without explicit constants) is found in Pisier [66, Lem 4.14].

**Lemma A.9.** *Let $\mathbf{X} = (X_1, \dots, X_m)$ be a standard Gaussian vector on $\mathbb{R}^m$.*

*(a) For $1 \le q < \infty$ we have*

$$\sqrt{\frac{2}{\pi}}\,m^{1/q} \le \mathbb{E}\,\|\mathbf{X}\|_q \le K\sqrt{q}\,m^{1/q}\,.$$

*(b) There exist constants $c, C > 0$ such that*

$$c\,\sqrt{1+\log m} \le \mathbb{E}\,\|\mathbf{X}\|_\infty \le C\,\sqrt{1+\log m}\,.$$

*Proof.* (a) The upper bound follows from

$$\mathbb{E}\,\|\mathbf{X}\|_q = \mathbb{E}\left(\sum_{j=1}^m |X_j|^q\right)^{1/q} \overset{\text{Jensen}}{\le} \left(\sum_{j=1}^m \mathbb{E}\,|X_j|^q\right)^{1/q},$$

and Lemma A.8.

For the lower bound we use the triangle inequality with the vector of absolute values $\operatorname{abs}\mathbf{X} := (|X_1|, \dots, |X_m|)$,

$$\mathbb{E}\,\|\mathbf{X}\|_q \ge \|\,\mathbb{E}\operatorname{abs}\mathbf{X}\|_q = \left(\sum_{j=1}^m (\mathbb{E}\,|X_j|)^q\right)^{1/q} = \sqrt{\frac{2}{\pi}}\,m^{1/q}\,.$$

(b) The upper bound is obtained by a comparison to some $q$-norm for $q \ge 1$, from (a) we have

$$\mathbb{E}\,\|\mathbf{X}\|_\infty \le \mathbb{E}\,\|\mathbf{X}\|_q \overset{\text{Jensen}}{\le} \left(\mathbb{E}\,\|\mathbf{X}\|_q^q\right)^{1/q} \le K\,\sqrt{q}\,m^{1/q}\,.$$

For the choice $q = 1 + \log m$ we have $m^{1/q} = \exp\left(\frac{\log m}{1+\log m}\right) < \mathrm{e}$, so that we obtain the desired upper bound with $C = K\mathrm{e} \approx 2.18884$.

Finally the lower bound. Using Lemma A.7 with $x = \mathrm{e}\,m$, we get

$$\mathbb{P}\left\{|X_j| > \alpha\sqrt{1+\log m}\right\} \ge \frac{1}{\mathrm{e}\,m}\,.$$

By this,

$$\mathbb{P}\Big\{\max_{j=1,\dots,m} |X_j| \le \alpha\sqrt{1+\log m}\Big\} \le \left(1 - \frac{1}{\mathrm{e}\,m}\right)^m \le \exp\left(-\mathrm{e}^{-1}\right).$$

We conclude

$$\begin{aligned}\mathbb{E}\,\|\mathbf{X}\|_\infty = \mathbb{E}\max_{j=1,\dots,m} |X_j| &\ge \alpha\sqrt{1+\log m}\,\left(1 - \mathbb{P}\Big\{\max_{j=1,\dots,m} |X_j| \le \alpha\sqrt{1+\log m}\Big\}\right)\\ &\ge \alpha\left(1 - \exp\left(-\mathrm{e}^{-1}\right)\right)\,\sqrt{1+\log m}\,.\end{aligned}$$

This gives us $c = \alpha\,(1 - \exp(-\mathrm{e}^{-1})) \approx 0.277159$. □

# Appendix B

# Useful Inequalities

## B.1 Combinatorics

The following two lemmas are well known.

**Lemma B.1** (Sum of binomial coefficients)**.** *For $k = 1, \dots, d$ we have*

$$\sum_{l=0}^{k} \binom{d}{l} < \left(\frac{\mathrm{e}\, d}{k}\right)^k .$$

*Proof.* For $k = 1$ we have

$$\mathrm{LHS} = 1 + d < \mathrm{e}\, d = \mathrm{RHS}\,.$$

For $k = d$ we have

$$\mathrm{LHS} = 2^d < \mathrm{e}^d = \mathrm{RHS}\,.$$

This completes the cases $d = 1, 2$.

Assume that the lemma holds for $d$ and $k = 1, \dots, d$. We show that it holds for $d + 1$ and $k = 2, \dots, d$ as well.

$$\begin{aligned}
\sum_{l=0}^{k} \binom{d+1}{l} &= \sum_{l=0}^{k} \binom{d}{l} + \sum_{l=0}^{k-1} \binom{d}{l} \\
&< \left(\frac{\mathrm{e}\, d}{k-1}\right)^{k-1} + \left(\frac{\mathrm{e}\, d}{k}\right)^k \\
&= \left(\frac{\mathrm{e}}{k}\right)^k \left( \underbrace{\frac{\left(1 + \frac{1}{k-1}\right)^{k-1}}{\mathrm{e}}}_{<1} k\, d^{k-1} + d^k \right) \\
&< \left(\frac{\mathrm{e}\,(d+1)}{k}\right)^k .
\end{aligned}$$

By induction, this completes the proof. □

**Lemma B.2** (Central binomial coefficient)**.** *For $d > 0$ we have*

$$\frac{1}{2}\frac{2^d}{\sqrt{d}} \leq \binom{d}{\lfloor d/2 \rfloor} \leq \sqrt{\frac{2}{\pi}}\frac{2^d}{\sqrt{d}}\,.$$

*Proof.* For odd $d = 2k+1$, $k \in \mathbb{N}_0$, by induction we see that $\binom{2k+1}{k} 2^{-(2k+1)}\sqrt{2k+1}$ is monotonously increasing in $k$. Similarly, for even $d = 2k$, $k \in \mathbb{N}$, we observe that $\binom{2k}{k} 2^{-2k}\sqrt{2k}$ is increasing. Hence the lower bound is done by considering $d = 1, 2$. For the upper bound we need the limit as $k \to \infty$, this can be obtained by Stirling's formula. □

## B.2 Quantitative Central Limit Theorem

**Proposition B.3** (Berry-Esseen inequality)**.** *Let $X_1, X_2, \ldots$ be iid random variables with zero mean, unit variance and finite third absolute moment $\beta_3$. Then there exists a universal constant $C_0$ such that*

$$\left|\mathbb{P}\left\{\frac{1}{\sqrt{d}}\sum_{j=1}^{d} X_j \leq x\right\} - \Phi(x)\right| \leq \frac{C_0\,\beta_3}{\sqrt{d}}\,,$$

*where $\Phi(\cdot)$ is the cumulative distribution function of the univariate standard normal distribution.*
*The best known estimates on $C_0$ are*

$$C_E := \frac{\sqrt{10}+3}{6\sqrt{2\pi}} = 0.409732\ldots \leq C_0 < 0.4748$$

*see Shevtsova [70].*

**Corollary B.4.** *Let $a := \lceil \frac{d}{2} + \alpha\frac{\sqrt{d}}{2} \rceil$ and $b := \lfloor \frac{d}{2} + \beta\frac{\sqrt{d}}{2} \rfloor$ with real numbers $\alpha < \beta$. Then we have*

$$\frac{1}{2^d}\sum_{k=a}^{b}\binom{d}{k} \;\geq\; \Phi(\beta) - \Phi(\alpha) - \frac{2\,C_0}{\sqrt{d}}\,.$$

*Proof.* Let $X_1, \ldots, X_d \overset{\text{iid}}{\sim} \text{unif}\{0,1\}$ be Bernoulli random variables and $Z_j := 2\,X_j - 1$ the corresponding Rademacher random variables. Note that the $Z_i$ have zero mean, unit variance, and third absolute moment $\beta_3 = 1$. Applying Proposition B.3 twice to the $Z_j$, we obtain

$$\begin{aligned}
\frac{1}{2^d}\sum_{k=a}^{b}\binom{d}{k} &= \mathbb{P}\left\{\frac{d}{2} + \alpha\frac{\sqrt{d}}{2} \leq \sum_{j=1}^{d} X_j \leq \frac{d}{2} + \beta\frac{\sqrt{d}}{2}\right\} \\
&= \mathbb{P}\left\{\alpha \leq \frac{1}{\sqrt{d}}\sum_{j=1}^{d} Z_j \leq \beta\right\} \\
&\geq \Phi(\beta) - \Phi(\alpha) - \frac{2\,C_0}{\sqrt{d}}\,.
\end{aligned}$$

□

# Abbreviations and Symbols

### Abbreviations

| | |
|---|---|
| a.s. | almost surely |
| IBC | information-based complexity |
| iid | independent and identically distributed (random variables) |
| LHS/RHS | *left-hand side*/*right-hand side* (of an equation or inequality referred to) |
| RKHS | reproducing kernel Hilbert space |

### On Functions and Real Numbers

| | |
|---|---|
| $a_+ := \max\{a, 0\}$ | positive part of a real number $a \in \mathbb{R}$ |
| $\lfloor a \rfloor$ | $a \in \mathbb{R}$ rounded down to an integer |
| $\lceil a \rceil$ | $a \in \mathbb{R}$ rounded up to an integer |
| $\mathbb{1}[\mathit{Statement}]$ | characteristic function, yielding 1 if *Statement* is true, and taking the value 0 if *Statement* is false |
| $\delta_{ij} = \mathbb{1}[i = j]$ | Kronecker symbol |
| log | natural logarithm |
| $f \preceq g$ | $f$ and $g$ non-negative functions on a common domain and $f \leq C\, g$ for a constant $C > 0$ |
| $f \asymp g$ | $f \preceq g$ and $g \preceq f$, that is, $c\, g \leq f \leq C\, g$ with $c, C > 0$ |
| $a_n \prec b_n$ for $n \to \infty$ | $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}} \subset [0, \infty)$ with $\frac{a_n}{b_n} \xrightarrow[n\to\infty]{} 0$ |
| $a_n \preceq b_n$ for $n \to \infty$ | there exist $n_0 \in \mathbb{N}$ and $C > 0$ such that for $n \geq n_0$ we have $a_n \leq C\, b_n$ |

### Vectors and Normed Spaces

| | |
|---|---|
| $\mathbf{x} = (x_1, \ldots, x_m)$ | vector in $\mathbb{R}^m$ (or $\ell_p$) with entries $\mathbf{x}(i) = x_i$ |
| $\mathbf{x}_A = (x_i)_{i\in A} \in \mathbb{R}^A$ | sub-vector for $A \subseteq \{1, \ldots, m\}$ |
| $[k] := \{1, \ldots, k\}$ | first $k$ natural numbers |
| $\mathbf{0} := (0, \ldots, 0)$ | vector with all the entries set to 0 |
| $\mathbf{1} := (1, \ldots, 1)$ | vector with all the entries set to 1 |
| $[\![\mathbf{x}, \mathbf{z}]\!] := \bigtimes_{j=1}^{m} [x_j, z_j]$ | closed cuboid with vector-valued interval boundaries $\mathbf{x} \leq \mathbf{z}$ |
| $\langle \mathbf{x}, \mathbf{z} \rangle$ | scalar product for vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^m = \ell_2^m$ |

| | |
|---|---|
| $\ell_p^m$ | $\mathbb{R}^m$ with the norm $\|\mathbf{x}\|_p := (\sum_{i=1}^m \|x_i\|^p)^{\frac{1}{p}}$ for $1 \leq p < \infty$, or $\|\mathbf{x}\|_\infty := \max \|x_i\|$ for $p = \infty$, analogously for $\ell_p = \mathbb{R}^\mathbb{N}$ |
| $\|\mathbf{x}\|_p$ | in some contextes instead of $\|\mathbf{x}\|_p$ for the $\ell_p^d$-norm |
| $\|\mathbf{x}\|_0$ | number of non-zero entries of $\mathbf{x} \in \mathbb{R}^d$ |
| $B_p^m$ | unit ball within $\ell_p^m$ |
| $\mathbf{e}_i$ | standard basis vector in $\mathbb{R}^m = \ell_p^m$, or $\mathbb{R}^\mathbb{N}$ |
| $L_p(\rho)$ | for $(Q, \Sigma, \rho)$ being a measure space, $L_p(\rho)$ is the space of measureable functions $f : Q \to \mathbb{R}$ that are bounded in the norm $\|f\|_p := \left(\int \|f\|^p \,\mathrm{d}\rho\right)^{1/p}$ for $1 \leq p < \infty$, or $\|f\|_\infty := \operatorname{ess\,sup}_{\mathbf{x}\in Q} \|f(\mathbf{x})\|$ for $p = \infty$; more precicely it is the space of equivalence classes of functions that are indistinguishable with respect to the metric induced by the norm |
| $L_p(Q)$ | $L_p$ on a domain $Q \subseteq \mathbb{R}^d$ with the $d$-dimensional Lebesgue measure $\lambda^d = \mathrm{Vol}$ |

**Operators**

| | |
|---|---|
| $\mathcal{L}(\mathbb{R}^m)$ | set of linear operators $\mathbb{R}^m \to \mathbb{R}^m$ |
| $\|J\|_{2\to F}$ | operator norm of a linear operator $J : \ell_2^m \to \widetilde{F}$ between normed spaces |
| $\operatorname{rk} T$ | rank of a linear operator $T$ |
| $\operatorname{tr} P$ | trace of a linear operator $P \in \mathcal{L}(\mathbb{R}^m)$ |
| $\operatorname{img} J$ | image $J(\mathbb{R}^m)$ of an operator $J : \mathbb{R}^m \to \widetilde{F}$ |
| $\widetilde{F} \hookrightarrow G$ | identity mapping, $\widetilde{F}$ is identified with a subset of $G$ |

**Analysis and Topology**

| | |
|---|---|
| $\nabla_\mathbf{v} f$ | directional derivative $[\nabla_\mathbf{v} f](\mathbf{x}) := \lim_{h\to 0} \frac{f(\mathbf{x}+h\mathbf{v})-f(\mathbf{x})}{h}$ of a $d$-variate function $f$ |
| $\partial_i f$ | partial derivative of a $d$-variate function $f$, that is $\nabla_{\mathbf{e}_i} f$, into the direction $\mathbf{e}_i$ of the $i$-th coordinate |

**Stochastics and Measure Theory**

| | |
|---|---|
| $(\Omega, \Sigma, \mathbb{P})$ | suitable probability space |
| $\mathbb{E}$ | expectation, i.e. the integration over $\omega \in \Omega$ with respect to $\mathbb{P}$ |
| $\Phi(x)$ | cummulative distribution function of a standard normal variable |
| $\mathbf{X} = (X_1, \ldots, X_m)$ | random vector |
| $\mathrm{unif}(A)$ | uniform distribution on a finite set $A$ |
| $\#A$ | number of elements of a set $A$ |

# Bibliography


[1] R. J. Adler. An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. *IMS Lecture Notes-Monograph Series*, 12, 1990. 60, 63, 64, 65, 66, 72, 74, 86, 87, 129

[2] R. J. Adler, J. E. Taylor. *Random Fields and Geometry.* Springer, 2007. 130

[3] G. Alberti, L. Ambrosio. A geometrical approach to monotone functions in $\mathbb{R}^n$. *Mathematische Zeitschrift*, 230:259–316, 1999. 90

[4] K. Amano, A. Maruoka. On learning monotone Boolean functions under the uniform distribution. *Theoretical Computer Science*, 350:3–12, 2006. 91

[5] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. 60

[6] N. S. Bakhvalov. On the approximate calculation of multiple integrals. *Vestnik MGU, Ser. Math. Mech. Astron. Phys. Chem.* 4:3–18, 1959, in Russian. English translation: *Journal of Complexity*, 31(4):502–516, 2015. 5, 45

[7] R. Bellman. *Dynamic Programming.* Princeton University Press, 1957. 7

[8] A. Blum, C. Burch, J. Langford. On learning monotone Boolean functions. *Proceedings of the 39th Annual Symposium on Foundations of Computer Science*, 408–415, 1998. v, xi, 91, 98, 99, 103, 105, 109, 112

[9] N. H. Bshouty, C. Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747-770, 1996. v, xi, 3, 6, 15, 91, 100, 114, 116, 117

[10] B. Carl. Entropy numbers, $s$-numbers, and eigenvalue problems. *Journal of Functional Analysis*, 41:290–306, 1981. 15

[11] K. Chandrasekaran, A. Deshpande, S. Vempala. Sampling $s$-concave functions: the limit of convexity based isoperimetry. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. Lecture Notes in Computer Science*, 5687:420–433, 2009. 90

[12] F. Cobos, T. Kühn, W. Sickel. Optimal approximation of multivariate periodic Sobolev functions in the sup-norm. *Journal of Functional Analysis*, 270(11):4196–4212, 2016. iv, x, 58, 66, 68, 76, 77, 78, 79, 87

[13] R. M. Dudley. Sample functions of the Gaussian process. *The Annals of Probability*, 1(1):66–103, 1973. 65

[14] L. C. Evans. *Partial Differential Equations*. Graduate Studies in Mathematics, 19, American Mathematical Society, 1998. 53

[15] G. Fang, L. Duan. The information-based complexity of approximation problem by adaptive Monte Carlo methods. *Science in China Series A: Mathematics*, 51(9):1679–1689, 2008. 54, 55, 86

[16] X. Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes, Ecole d'Eté de Probabilités de Saint-Flour IV-1974. *Lecture Notes in Mathematics*, 480:1–96, 1975. 64

[17] S. Foucart, H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013. 38, 50, 52

[18] F. Gao, J. A. Wellner. Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98:1751–1764, 2007. 90

[19] M. R. Garey, D. S. Johnson. *Computers and Intractability – A Guide to the Theory of NP-Completeness*. A Series of Books in the Mathematical Sciences, 1979. 8

[20] A. Garnaev, E. Gluskin. On widths of a Euclidean ball. *Soviet Mathematics Doklady*, 30:200–204, 1984, English translation. 38

[21] M. Gnewuch, H. Woźniakowski. Quasi-polynomial tractability. *Journal of Complexity*, 27(3–4):312–330, 2011. 8

[22] S. Heinrich. Lower bounds for the complexity of Monte Carlo function approximation. *Journal of Complexity*, 8(3):277–300, 1992. ii, viii, 10, 14, 15, 16, 17, 18, 20, 24, 25, 27, 31, 37, 49, 52, 53, 54, 55, 129

[23] S. Heinrich. Randomized approximation of Sobolev embeddings. *Proceedings of the MCQMC 2006*, 445–459, 2008. 56

[24] S. Heinrich. Randomized approximation of Sobolev embeddings, II. *Journal of Complexity*, 25(5):455–472, 2009. 56

[25] S. Heinrich. Randomized approximation of Sobolev embeddings, III. *Journal of Complexity*, 25(5):473–507, 2009. 56

[26] S. Heinrich, B. Milla. The randomized complexity of indefinite integration. *Journal of Complexity*, 27(3–4):352–382, 2011. 13, 14, 60, 126

[27] A. Hinrichs, E. Novak, M. Ullrich, H. Woźniakowski. Product rules are optimal for numerical integration in classical smoothness spaces. *Journal of Complexity*, 38:39–49, 2017. 46

[28] A. Hinrichs, E. Novak, H. Woźniakowski. The curse of dimensionality for the class of monotone functions and for the class of convex functions. *Journal of Approximation Theory*, 163(8):955–965, 2011. v, xi, 89, 90, 91, 95

[29] C. V. Hutton, J. S. Morrell, J. R. Retherford. Diagonal operators, approximation numbers, and Kolmogoroff diameters. *Journal of Approximation Theory*, 16(1):48–80, 1976. 15, 72

[30] B. Kashin. Diameters of some finite-dimensional sets and classes of smooth functions. *Mathematics of the USSR Izvestija*, 11(2):317–333, 1977, English Translation. 38

[31] B. Kashin. Widths of Sobolev classes of small-order smoothness. *Vestnik Moskovskogo Universiteta Seriya 1 Matematika Mekhanika*, 5:50–54, 1981, in Russian. 50

[32] C. Katscher, E. Novak, K. Petras. Quadrature formulas for multivariate convex functions. *Journal of Complexity*, 12(1):5–16, 1996. 90

[33] M. Kearns, M. Li, L. Valiant. Learning Boolean formulas. *Journal of the Association for Computing Machinery*, 41(6):1298–1328, 1994. 91

[34] M. Kearns, L. Valiant. Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM*, 41(1):67–95, 1994. 91

[35] H. König. *Eigenvalue Distribution of Compact Operators*. Birkhäuser, 1986. 55

[36] K. A. Kopotun. Approximation of $k$-monotone functions. *Journal of Approximation Theory*, 94:481–493, 1998. 90

[37] A. D. Korshunov. Monotone Boolean functions. *Russian Mathematical Surveys*, 58(5):929–1001, 2003. 90, 100

[38] S. N. Kudryavtsev. Bernstein width of a class of functions of finite smoothness. *Sbornik Mathematics*, 190(4):539–560, 1999. 16, 36

[39] R. J. Kunsch. Bernstein numbers and lower bounds for the Monte Carlo error. *Proceedings of the MCQMC 2014*, 471–488, 2016. iii, viii, 17, 18, 37, 38, 40

[40] F. Y. Kuo, G. W. Wasilkowski, H. Woźniakowski. Multivariate $L_\infty$ approximation in the worst case setting over reproducing kernel Hilbert spaces *Journal of Approximation Theory*, 152:135–160, 2008. iv, x, 58, 66, 68, 69, 77, 87

[41] M. Ledoux, M. Talagrand. *Probability in Banach Spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete, 3. Folge, Band 23, Springer-Verlag, 1991. 63, 65

[42] D. Lewis. Ellipsoids defined by Banach ideal norms. *Mathematika*, 26:18–29, 1979. 26

[43] M. A. Lifshits. *Gaussian Random Functions*. Mathematics and Its Applications, Kluwer Academic Publishers, 1995. 63, 64, 65, 129

[44] V. E. Maiorov. Discretization of the problem of diameters. *Uspekhi Matematicheskikh Nauk*, 30(6):179–180, 1975, in Russian. 54

[45] M. B. Marcus, G. Pisier. *Random Fourier Series with Applications to Harmonic Analysis*. Annals of Mathematics Studies, Princeton University Press, 1981. 86, 87

[46] P. Mathé. Random approximation of Sobolev embeddings. *Journal of Complexity*, 7(3):261–281, 1991. iii, ix, 15, 20, 27, 45, 47, 49, 54

[47] P. Mathé. A minimax principle for the optimal error of Monte Carlo methods. *Constructive Approximation*, 9:23-39, 1993. 6, 51, 102

[48] M. Meyer, A. Pajor. Sections of the unit ball of $\ell_p^n$. *Journal of Functional Analysis*, 80(1):109–123, 1988. 38

[49] V. K. Nguyen. Bernstein numbers of embeddings of isotropic and dominating mixed Besov spaces. *Mathematische Nachrichten*, 288(14–15):1694–1717, 2015. 16, 36

[50] V. K. Nguyen. Weyl and Bernstein numbers of embeddings of Sobolev spaces with dominating mixed smoothness. *Journal of Complexity*, 36:46–73, 2016. 16, 36

[51] E. Novak. Deterministic and Stochastic Error Bounds in Numerical Analysis. *Lecture Notes in Mathematics*, 1349, 1988. 45

[52] E. Novak. Optimal linear randomized methods for linear operators in Hilbert spaces. *Journal of Complexity*, 8(1):22–36, 1992. 17

[53] E. Novak. Quadrature formulas for monotone functions. *Proceedings of the American Mathematical Society*, 115(1):59–68, 1992. 6, 90, 92

[54] E. Novak, K. Petras. Optimal stochastic quadrature formulas for convex functions. *BIT Numerical Mathematics*, 34(2):288–294, 1994. 90

[55] E. Novak, H. Woźniakowski. *Tractability of Multivariate Problems*, Volume I, Linear Information. European Mathematical Society, 2008. 6, 8, 13, 62, 76, 83, 88, 100, 125

[56] E. Novak, H. Woźniakowski. Approximation of infinitely differentiable multivariate functions is intractable. *Journal of Complexity*, 25(4):398–404, 2009. iii, ix, 41, 44

[57] E. Novak, H. Woźniakowski. *Tractability of Multivariate Problems*, Volume II, Standard Information for Functionals. European Mathematical Society, 2010. 60, 62, 76, 88

[58] E. Novak, H. Woźniakowski. *Tractability of Multivariate Problems*, Volume III, Standard Information for Operators. European Mathematical Society, 2012. 60, 62, 76

[59] R. O'Donnell, R. A. Servedio. Learning monotone decision trees in polynomial time. *SIAM Journal on Computing*, 37(3):827–844, 2007. 91

[60] K. Y. Osipenko. *Optimal Recovery of Analytic Functions.* Nova Science Publishers, 2000. 72

[61] K. Y. Osipenko, O. G. Parfenov. Ismagilov type theorems for linear, Gel'fand and Bernstein n-widths. *Journal of Complexity*, 11(4):474–492, 1995. 16, 58, 66, 68, 71

[62] A. Papageorgiou. Integration of monotone functions of several variables. *Journal of Complexity*, 9(2):252–268, 1993. 90, 92

[63] A. Pietsch. $s$-Numbers of operators in Banach spaces. *Studia Mathematica*, 51(3):201–223, 1974. 15, 16, 49, 67

[64] A. Pinkus. *n-Widths in Approximation Theory.* Springer-Verlag, 1985. 38, 54

[65] G. Pisier. Probabilistic Methods in the Geometry of Banach Spaces. *Lecture Notes in Mathematics*, 1206, 1985. 129, 130

[66] G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry.* Cambridge University Press, 1989. 15, 18, 26, 27, 130, 132

[67] L. Pitt, L. G. Valiant. Computational limitations on learning from examples. *Journal of the Association for Computing Machinery*, 35(4):965–984, 1988. 91

[68] K. Ritter. Average-Case Analysis of Numerical Problems. *Lecture Notes in Mathematics*, 1733, 2000. 9, 60, 62, 76

[69] H. Robbins, S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. 90

[70] I. Shevtsova. On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. Available on *arXiv*, 2001. 135

[71] P. Siedlecki, M. Weimar. Notes on $(s,t)$-weak tractability: A refined classification of problems with (sub)exponential information complexity. *Journal of Approximation Theory*, 200:227–258, 2015. 8

[72] S. A. Smolyak. *On Optimal Restoration of Functions and Functionals of Them.* Candidate dissertation, Moscow State University, 1965, in Russian. 51, 72

[73] J. F. Traub, G. W. Wasilkowski, H. Woźniakowski. *Information-Based Complexity.* Academic Press, 1988. 1, 4, 15, 20, 49, 125

[74] H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators.* Barth, 1995. 53

[75] J. Vybíral. Widths of embeddings in function spaces. *Journal of Complexity*, 24:545–570, 2008. 54

[76] J. Vybíral. Average best $m$-term approximation. *Constructive Approximation*, 36:83-115, 2012. 40

[77] J. Vybíral. Weak and quasi-polynomial tractability of approximation of infinitely differentiable functions. *Journal of Complexity*, 30(2):48–55, 2014. iii, ix, 43

[78] G. Wahba. Spline Models for Observational Data. *CBMS-NSF Regional Conference Series in Applied Mathematics*, 1990. 60

[79] M. Weimar. Tractability results for weighted Banach spaces of smooth functions. *Journal of Complexity*, 28(1):59–75, 2012. 44

[80] G. Xu. On weak tractability of the Smolyak algorithm for approximation problems. *Journal of Approximation Theory*, 192:347–361, 2015. 44

# Ehrenwörtliche Erklärung

Hiermit erkläre ich,

- dass mir die Promotionsordnung der Fakultät bekannt ist,
- dass ich die Dissertation selbst angefertigt habe, keine Textabschnitte oder Ergebnisse eines Dritten oder eigenen Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönliche Mitteilungen und Quellen in meiner Arbeit angegeben habe,
- dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,
- dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts wurde ich durch Prof. Dr. Erich Novak unterstützt.

Ich habe weder die gleiche, noch eine in wesentlichen Teilen ähnliche oder andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht.

Jena, 31. März 2017 Robert Kunsch