Convergence Guarantees
for Adaptive Bayesian Quadrature Methods

Motonobu Kanagawa
University of Tübingen
Maria-von-Linden-Strasse 6, 72076 Tübingen.
motonobu.kanagawa@uni-tuebingen.de

Philipp Hennig
University of Tübingen and Max Planck Institute for Intelligent Systems
Maria-von-Linden-Strasse 6, 72076 Tübingen.
philipp.hennig@uni-tuebingen.de

May 27, 2019

Abstract

Adaptive Bayesian quadrature (ABQ) is a powerful approach to numerical integration that empirically
does better than Monte Carlo integration on problems of medium dimensionality (where non-adaptive quadrature is not competitive). Its key ingredient is an acquisition function that changes as a function of previously collected values of the integrand. While this adaptivity appears to be empirically powerful, it complicates analysis. Consequently, there are no theoretical guarantees so far for this class of methods. In this work, for a broad class of adaptive Bayesian quadrature methods, we prove consistency, deriving non-tight but informative convergence rates. To do so we introduce a new concept we call weak adaptivity. In guaranteeing consistency of ABQ, weak adaptivity is notionally similar to the ideas of detailed balance and ergodicity in Markov Chain Monte Carlo methods, which allow sufficient conditions for consistency of MCMC. Likewise, our results identify a large and flexible class of adaptive Bayesian quadrature rules as consistent, within which practitioners can develop empirically efficient methods.

1 Introduction

Numerical integration, or quadrature/cubature, is a fundamental task in many areas of science and engineering. This includes machine learning and statistics, where such problems arise when computing marginals and conditionals in probabilistic inference problems. In particular in hierarchical Bayesian inference, quadrature is generally required for the computation of the marginal likelihood, the key quantity for model selection, and for prediction, for which latent variables are to be marginalized out.

To describe the problem, let \( \Omega \) be a compact metric space (such as a bounded and closed subset in \( \mathbb{R}^d \)), \( \pi : \Omega \to \mathbb{R} \) be a known density function, and \( f : \Omega \to \mathbb{R} \) be an integrand, a known function such that the function value \( f(x) \in \mathbb{R} \) can be obtained for any given query \( x \in \Omega \). The task of quadrature is to numerically compute the integral (assumed to be intractable analytically)

\[
\int f(x)\pi(x)dx.
\]
This is done by evaluating the function values $f(x_1), \ldots, f(x_n)$ at design points $x_1, \ldots, x_n \in \Omega$ and using them to approximate $f$ and the integral. The points $x_1, \ldots, x_n$ should be “good” in the sense that $f(x_1), \ldots, f(x_n)$ provide useful information for computing the integral.

Monte Carlo methods are the classic alternative, where $x_1, \ldots, x_n$ are randomly generated from a proposal distribution and the integral is approximated as $\sum_{i=1}^{n} w_i f(x_i)$, with $w_1, \ldots, w_n$ being importance weights. Monte Carlo is consistent, but may make inefficient use of computational resources [e.g. 23], requiring a huge number of points to give an accurate estimate of the integral. The inefficiency of Monte Carlo stems partly from the fact that it covers a wide class of functions, such as that of bounded continuous functions. Its focus is to obtain sample points that approximate well the distribution of $\pi$ in a weak sense. If, on the contrary, the purpose is to compute the integral for one specific target integrand $f$, then there may be room for drastic improvement by adaptively generating informative design points. For instance, if the function value $f(x)$ changes more rapidly in a certain region than the rest of the domain, it would be more efficient to allocate a larger number of points in that region, rather than uniformly allocating design points in the entire domain.

Adaptive Bayesian quadrature (ABQ) is a recent approach from machine learning that actively, sequentially, and deterministically selects design points to adapt to the target integrand [26, 27, 13, 1, 8]. It is an extension of Bayesian quadrature (BQ) [25, 12, 7, 18], a probabilistic numerical method for quadrature that makes use of prior knowledge about the integrand, such as smoothness and structure, via a Gaussian process (GP) prior. A drawback of vanilla BQ is that the Gaussian process model prevents the use of certain kinds of relevant knowledge about the integrand, such as it being positive (or non-negative), because they can not be encoded in a Gaussian distribution. Positive integrands are ubiquitous in machine learning and statistics, where integration tasks emerge in the marginalization and conditioning of probability density functions, which are positive by definition. In ABQ such prior knowledge is modelled by describing integrand as given by a certain transformation (or warping) of a GP — for instance, an exponentiated GP [27, 26, 8] or a squared GP [13]. ABQ methods with such transformations have empirically been shown to improve upon both standard BQ and Monte Carlo, leading to state-of-the-art wall-clock time performance on problems of medium dimensionality.

If the transformation is nonlinear, as in the examples above, the transformed GP no longer allows an analytic expression for its posterior process, and thus approximations are used to obtain a tractable acquisition function. In contrast to the posterior covariance of GPs, these acquisition functions then become dependent on previous observations, making the algorithm adaptive. This twist seems to be critical for ABQ methods’ superior empirical performance, but it complicates analysis. Thus, there has been no theoretical guarantee for their convergence, rendering them heuristics in practice. This is problematic since integration is usually an intermediate computational step in a larger system, and thus must be reliable. This paper provides the first convergence analysis for ABQ methods.

In Sec. 2 we review ABQ methods, and formulate a generic class of acquisition functions that cover those of [13, 1, 2, 8]. Our convergence analysis is done for this class. We also derive an upper-bound on the quadrature error using a transformed integrand, which is applicable to any design points and given in terms of the GP posterior variance (Prop. 2.1). In Sec. 3 we establish a connection between ABQ and certain weak greedy algorithms (Thm. 3.3). This is based on a new result that the scaled GP posterior variance can be interpreted in terms of a certain projection in a Hilbert space (Lemma 3.1). Using this connection, we derive convergence rates of ABQ methods in Sec. 4. We provide sufficient conditions for ABQ methods to converge.

To motivate the relevance of our results, consider the parallel situation in Monte Carlo methods: Exact sampling is also hard and inefficient, so Markov Chain Monte Carlo methods are a popular approximate alternative. Not every MCMC method is consistent though. The notions of detailed balance and ergodicity (which go back to Boltzmann) are the crucial tool [cf. §6.5 & 6.6 in 30]. MCMC chains that satisfy these properties (and a few other technical constraints) are consistent. It is comparably straightforward to design such Markov Chains, so researchers have been able to build an ecosystem of efficient MCMC methods within it. Typically such methods have little additional theoretical guarantees and are mostly motivated empirically

\footnote{To give an example, Wenliang et al. [30 Fig. 3] used $10^{10}$ Monte Carlo samples to estimate the the normalizing constant of their model.}
and intuitively. Analogously, we here develop a relatively general notion for active exploration that we term *weak adaptivity*. An ABQ method that satisfies weak adaptivity (and a few additional technical constraints) is consistent, and the conceptual space of weakly adaptive ABQ methods is large and flexible. We hope that our results spark a practical interest in the design of empirically efficient acquisition functions, to extend the reach of quadrature to problems of higher and higher dimensionalities.

**Related Work.** For standard BQ methods, and the corresponding *kernel quadrature* rules, convergence properties have been studied extensively [e.g. 6 16 3 37 18 9 7 23 17]. In particular, there are methods that deterministically generate design points [10 4 14 6 9]. These methods are *not* adaptive, as design points are generated independently to the function values of the target integrand.

Our analysis is technically related to the work by Santin and Haasdonk [31] on solvers for partial differential equations, which analyzed the so-called *P-greedy algorithm*, an algorithm to sequentially obtain design points using the GP posterior variance as an acquisition function. Our results can be regarded as a generalization of their result so that the acquisition function can include i) a scaling and a transformation of the GP posterior variance and ii) a data-dependent term that takes care of adaptation; see [4] for details.

Adaptive methods have also been theoretically studied in the information-based complexity literature [19–22]. The key result is that optimal points for quadrature can be obtained *without* observing actual function values, if the hypothesis class of functions is symmetric and convex (e.g. the unit ball in a Hilbert space): in this case adaptation does *not* help improve the performance. On the other hand, if the hypothesis class is either asymmetric or nonconvex, then adaptation may be helpful. For instance, a class of positive functions is asymmetric if only one of \( f \) or \( -f \) can be positive. These results thus support the choice of acquisition functions of existing ABQ methods, where the adaptivity to function values is motivated by modeling the positivity of the integrand.

**Notation.** \( \mathbb{N} \) denotes the set of positive integers, \( \mathbb{R} \) the real line, and \( \mathbb{R}^d \) the \( d \)-dimensional Euclidean space for \( d \in \mathbb{N} \). \( L_p(\Omega) \) for \( 1 \leq p < \infty \) is the Banach space of \( p \)-integrable functions, and \( L_\infty(\Omega) \) is that of essentially bounded functions.

## 2 Adaptive Bayesian Quadrature (ABQ)

We describe here ABQ methods, and present a generic form of acquisition functions that we analyze. We also derive an upper-bound on the quadrature error using a transformed integrand in terms of the GP posterior variance, motivating our analysis in the later sections. Throughout the paper we assume that the domain \( \Omega \) is a compact metric space.

### 2.1 Bayesian Quadrature with Transformation

ABQ methods deal with an integrand \( f \) that is a priori known to satisfy a certain constraint, for example \( f(x) > 0 \ \forall x \in \Omega \). Such a constraint is modeled by considering a certain transformation \( T: \mathbb{R} \to \mathbb{R} \), and assuming that there exists a latent function \( g: \Omega \to \mathbb{R} \) such that the integrand \( f \) is given as the transformation of \( g \), i.e., \( f(x) = T(g(x)), x \in \Omega \). Examples of \( T \) for modeling the positivity include i) the square transformation \( T(y) = \alpha + \frac{1}{2}y^2 \), where \( \alpha > 0 \) is a small constant such that \( \alpha > \inf_{x \in \Omega} f(x) > 0 \), assuming that \( f \) is bounded away from 0 [13]; and ii) the exponential transformation \( T(y) = \exp(y) \) [27 26 8]. Note that the identity map \( T(y) = y \) recovers standard Bayesian quadrature (BQ) methods [25 12 6 18]. To model the latent function \( g \), a Gaussian process (GP) prior [29] is placed over \( g \):

\[
g \sim \mathcal{GP}(m, k)
\]
where \( m : \Omega \to \mathbb{R} \) is a mean function and \( k : \Omega \times \Omega \) is a covariance kernel. Both \( m \) and \( k \) should be chosen to capture as much prior knowledge or belief about \( g \) (or its transformation \( f \)) as possible, such as smoothness and correlation structure; see e.g. [29 Chap. 4].

Assume that a set of points \( X_n := \{x_1, \ldots, x_n\} \subset \Omega \) are given, such that the kernel matrix \( K_n := (k(x_i, x_j))_{1 \leq i, j \leq n} \subset \mathbb{R} \) is invertible. Given the function values \( f(x_1), \ldots, f(x_n) \), define \( g_i(x) := z_i \in \mathbb{R} \) such that \( T(z_i) = f(x_i) \) for \( i = 1, \ldots, n \). Treating \( g(x_1), \ldots, g(x_n) \) as “observed data without noise,” the posterior distribution of \( g \) under the GP prior (1) is again given as a GP

\[
g(x_i, g(x_i))_{i=1}^n \sim \mathcal{GP}(m_{g,X_n}, k_{X_n}),
\]

where \( m_{g,X_n} : \Omega \to \mathbb{R} \) is the posterior mean function and \( k_{X_n} : \Omega \times \Omega \to \mathbb{R} \) is the posterior covariance kernel given by (see e.g. [29])

\[
m_{g,X_n}(x) := m(x) + k_n(x) \top K_n^{-1}(g_n - m_n),
\]
\[
k_{X_n}(x, x') := k(x, x') - k_n(x) \top K_n^{-1}k_n(x'),
\]

where \( k_n(x) := (k(x, x_1), \ldots, k(x, x_n))^\top \in \mathbb{R}^n, \ g_n := (g(x_1), \ldots, g(x_n))^\top \in \mathbb{R}^n \) and \( m_n = (m(x_1), \ldots, m(x_n))^\top \in \mathbb{R}^n \). Then a quadrature estimate\(^2\) for the integral \( \int f(x)\pi(x)dx \) is given as the integral \( \int T(m_{g,X_n}(x))\pi(x)dx \) of the transformed posterior mean function \( T(m_{g,X_n}) \). The posterior covariance for \( \int f(x)\pi(x)dx \) is given similarly; see [3] for details.

### 2.2 A Generic Form of Acquisition Functions

The key remaining question is how to select good design points \( x_1, \ldots, x_n \in \Omega \). ABQ methods sequentially and deterministically generate \( x_1, \ldots, x_n \) using an acquisition function. Many of the acquisition functions can be formulated in the generic form:

\[
x_{\ell+1} \in \arg\max_{x \in \Omega} a_\ell(x), \text{ where } a_\ell(x) = F(q^\ell(x)k_{X_\ell}(x))b_\ell(x), \quad (\ell = 0, 1, \ldots, n-1)
\]

where \( k_{X_\ell}(x, x) := k(x, x), F : [0, \infty) \to [0, \infty) \) is an increasing function such that \( F(0) = 0, q : \Omega \to (0, \infty) \) and \( b_\ell(x) : \Omega \to \mathbb{R} \) is a function that may change at each iteration \( \ell \), e.g., it may depend on the function values \( f(x_1), \ldots, f(x_\ell) \) of the target integrand \( f \). We analyse ABQ with this generic form [4], aiming for results with wide applicability. Here are some representative choices.

### Warped Sequential Active Bayesian Integration (WSABI) [13]:

Gunter et al. [13] employ the square transformation \( f(x) = T(g(x)) = \alpha + \frac{1}{2}g^2(x) \) with two acquisition functions: i) WSABI-L [13 Eq. 15], which is based on linearization of \( T \) and recovered with \( F(y) = y, q(x) = \pi(x) \) and \( b_\ell(x) = m_{g,X_\ell}(x) \); and ii) WSABI-M [13 Eq. 14], the one based on moment matching given by \( F(y) = y, q(x) = \pi(x) \) and \( b_\ell(x) = \frac{1}{2}k_{X_\ell}(x) + m_{g,X_\ell}^2 \).

### Moment-Matched Log-Transformation (MMLT) [8]:

Chai and Garnett [8 3rd raw in Table 1] use the exponential transformation \( f(x) = T(g(x)) = \exp(g(x)) \) with the acquisition function given by \( F(y) = \exp(y) - 1, q(x) = 1 \) and \( b_\ell(x) = \exp(k_{X_\ell}(x, x) + 2m_{g,X_\ell}(x)) \).

### Variational Bayesian Monte Carlo (VBMC) [11 2]:

Acerbi [2] uses the identity \( f(x) = T(g(x)) = g(x) \) with the acquisition function given by \( F(y) = y^\delta, q(x) = 1 \) and \( b_\ell(x) = \pi_\ell^2(x) \exp(\delta_3 m_{g,X_\ell}(x)) \), where \( \pi_\ell \) is the variational posterior at the \( \ell \)-th iteration and \( \delta_1, \delta_2, \delta_3 \geq 0 \) are constants: setting \( \delta_1 = \delta_2 = \delta_3 = 1 \) recovers the original acquisition function [11 Eq. 9]. Acerbi [11 Sec. 2.1] considers an integrand \( f \) that is defined as the logarithm of a joint density, while \( \pi \) is an intractable posterior that is gradually approximated by the variational posteriors \( \pi_\ell \).

---

\(^2\)The point is that, in contrast to the integral over \( f \), this estimate should be analytically tractable. This depends on the choices for \( T, k \) and \( \pi \). For instance, for \( T(y) = y \) or \( T(y) = \alpha + \frac{1}{2}y^2 \) with \( k \) and \( \pi \) Gaussian, the estimate can be obtained analytically [13], while for \( T(y) = \exp(y) \) one needs approximations; [cf. 8].
For the WSABI and MMLT, the acquisition function (4) is obtained by a certain approximation for the posterior variance of the integral \( \int f(x)\pi(x)dx = \int T(g(x))\pi(x)dx \); thus this is a form of uncertainty sampling. Such an approximation is needed because the posterior variance of the integral is not available in closed form, due to the nonlinear transformation \( T \). The resulting acquisition function includes the data-dependent term \( b_0(x) \), which encourages exploration in regions where the value of \( g(x) \) is expected to be large. This makes ABQ methods adaptive to the target integrand. Alas, it also complicates analysis. Thus there has been no convergence guarantee for these ABQ methods; which is what we aim to remedy in this paper.

### 2.3 Bounding the Quadrature Error with Transformation

Our first result, which may be of independent interest, is an upper-bound on the error for the quadrature estimate based on a transformation described in Sec. 2.1. It is applicable to any point set \( X_n = \{x_1, \ldots, x_n\} \), and the bound is given in terms of the posterior variance \( k_x(x, x) \). This gives us a motivation to study the behavior of this quantity for \( x_1, \ldots, x_n \) generated by ABQ (4) in the later sections. To state the result, we need to introduce the Reproducing Kernel Hilbert Space (RKHS) of the covariance kernel \( k \) of the GP prior. See e.g. [32, 33] for details of RKHS’s, and [5, 15] for discussions of their close but subtle relation to the GP notion.

Let \( H_k \) be the RKHS associated with the covariance kernel \( k \) of the GP prior (1), with \( \langle \cdot, \cdot \rangle_{H_k} \) and \( \| \cdot \|_{H_k} \) being its inner-product and norm, respectively. \( H_k \) is a Hilbert space consisting of functions on \( \Omega \), such that i) \( k(\cdot, \cdot) \in H_k \) for all \( x \in \Omega \), and ii) \( h(x) = \langle k(\cdot, x), h \rangle_{H_k} \) for all \( h \in H_k \) and \( x \in \Omega \) (the reproducing property), where \( k(\cdot, x) \) denotes the function of the first argument such that \( y \to k(y, x) \), with \( x \) being fixed. As a set of functions, \( H_k \) is given as the closure of the linear span of such functions \( k(\cdot, x) \), i.e., \( H_k = \text{span} \{k(\cdot, x) \mid x \in \Omega\} \), meaning that any \( h \in H_k \) can be written as \( h = \sum_{i=1}^{\infty} \alpha_i k(\cdot, x_i) \) for some \( (\alpha_i)_{i=1}^{\infty} \subset \mathbb{R} \) and \( (x_i)_{i=1}^{\infty} \subset \Omega \) such that \( \|h\|_{H_k}^2 = \sum_{i=1}^{\infty} \alpha_i^2 \|k(x_i, x_i)\| < \infty \). We are now ready state our assumption:

**Assumption 1.** \( T : \mathbb{R} \to \mathbb{R} \) is continuously differentiable. For \( f : \Omega \to \mathbb{R} \), there exists \( g : \Omega \to \mathbb{R} \) such that \( f(x) = T(g(x)), x \in \Omega \) and that \( \tilde{g} := g - m \in H_k \). It holds that \( \|k\|_{L_\infty(\Omega)} := \sup_{x \in \Omega} k(x, x) < \infty \) and \( \|m\|_{L_\infty(\Omega)} := \sup_{x \in \Omega} |m(x)| < \infty \).

The assumption \( \tilde{g} := g - m \in H_k \) is common in theoretical analysis of standard BQ methods, where \( T(g) = y \) and \( m = 0 \) [see e.g. [6, 37, 7] and references therein]. This assumption may be weakened by using proof techniques developed for standard BQ in the misspecified setting [16, 17], but we leave it for a future work. The other conditions on \( T, k \) and \( m \) are weak.

**Proposition 2.1.** (proof in Appendix A.1) Let \( \Omega \) be a compact metric space, \( X_n = \{x_1, \ldots, x_n\} \subset \Omega \) be such that the kernel matrix \( K_n = (k(x_i, x_j))_{i,j=1}^{n \times n} \in \mathbb{R}^{n \times n} \) is invertible, and \( \pi : \Omega \to [0, \infty) \) be such that \( \int_{\Omega} |\pi(x)|q(x)dx < \infty \). Suppose that Assumption 1 is satisfied. Then there exists a constant \( C_{\tilde{g}, m, k, T}\) depending only on \( \tilde{g}, m, k \) and \( T \) such that

\[
\left| \int f(x)\pi(x)dx - \int T(m_{g, X_n}(x))\pi(x)dx \right| \leq C_{\tilde{g}, m, k, T}C_{\pi/q}\sup_{x \in \Omega} q(x)\sqrt{k_{X_n}(x, x)}.
\]

Prop. 2.1 shows that to establish convergence guarantees for ABQ methods, it is sufficient to analyze the convergence behavior of the quantity \( \sup_{x \in \Omega} q(x)\sqrt{k_{X_n}(x, x)} \) for points \( X_n = \{x_1, \ldots, x_n\} \) generated from ABQ (4). This is what we focus on in the remainder.

### 3 Connections to Weak Greedy Algorithms in Hilbert Spaces

To analyze the quantity \( \sup_{x \in \Omega} q(x)\sqrt{k_{X_n}(x, x)} \) for points \( X_n = \{x_1, \ldots, x_n\} \) generated from ABQ (4), we show here that the ABQ can be interpreted as a certain weak greedy algorithm studied by DeVore et al. [11].
To describe this, let $\mathcal{H}$ be a (generic) Hilbert space and $\mathcal{C} \subset \mathcal{H}$ be a compact subset. To define some notation, let $h_1, \ldots, h_n \in \mathcal{C}$ be given. Denote by $S_n := \text{span}(h_1, \ldots, h_n) = \left\{ \sum_{i=1}^{n} \alpha_i h_i \mid \alpha_1, \ldots, \alpha_n \in \mathbb{R} \right\} \subset \mathcal{H}$ the linear subspace spanned by $h_1, \ldots, h_n$. For a given $h \in \mathcal{C}$, let $\text{dist}(h, S_n)$ be the distance between $h$ and $S_n$ defined by
\[
\text{dist}(h, S_n) := \inf_{g \in S_n} \|h - g\|_\mathcal{H} = \inf_{\alpha_1, \ldots, \alpha_n \in \mathbb{R}} \|h - \sum_{i=1}^{n} \alpha_i h_i\|_\mathcal{H},
\]
where $\|\cdot\|_\mathcal{H}$ denotes the norm of $\mathcal{H}$. Geometrically, this is the distance between $h$ and its orthogonal projection onto the subspace $S_n$. The task considered in [11] is to select $h_1, \ldots, h_n \in \mathcal{C}$ such that the worst case error in $\mathcal{C}$ defined by
\[
e_n(\mathcal{C}) := \sup_{h \in \mathcal{C}} \text{dist}(h, S_n)
\]
becomes as small as possible: $h_1, \ldots, h_n \in \mathcal{C}$ are to be chosen to approximate well the set $\mathcal{C}$.

The following weak greedy algorithm is considered in DeVore et al. [11]. Let $\gamma$ be a constant such that $0 < \gamma \leq 1$, and let $n \in \mathbb{N}$. First select $h_1 \in \mathcal{C}$ such that $\|h_1\|_\mathcal{H} \geq \gamma \sup_{h \in \mathcal{C}} \|h\|_\mathcal{H}$. For $\ell = 1, \ldots, n-1$, suppose that $h_1, \ldots, h_\ell$ have already been generated, and let $S_\ell = \text{span}(h_1, \ldots, h_\ell)$. Then select a next element $h_{n+1} \in \mathcal{C}$ such that
\[
\text{dist}(h_{n+1}, S_n) \geq \gamma \sup_{h \in \mathcal{C}} \text{dist}(h, S_\ell), \quad (\ell = 1, \ldots, n-1).
\]
In this paper we refer to such $h_1, \ldots, h_n$ as a $\gamma$-weak greedy approximation of $\mathcal{C}$ in $\mathcal{H}$ because, $\gamma = 1$ recovers the standard greedy algorithm, while $\gamma < 1$ weakens the “greediness” of this rule. DeVore et al. [11] derived convergence rates of the worst case error (5) as $n \to \infty$ for $h_1, \ldots, h_n$ generated from this weak greedy algorithm.

Weak Greedy Algorithms in the RKHS. To establish a connection to ABQ, we formulate the weak greedy algorithm in an RKHS. Let $\mathcal{H}_k$ be the RKHS of the covariance kernel $k$ as in Sec. 2.3 and $q(x)$ be the function in (4). We define a subset $\mathcal{C}_{k,q} \subset \mathcal{H}_k$ by
\[
\mathcal{C}_{k,q} := \{ q(x)k(\cdot, x) \mid x \in \Omega \} \subset \mathcal{H}_k.
\]
Note that $\mathcal{C}_{k,q}$ is the image of the mapping $x \to q(x)k(\cdot, x)$ with $\Omega$ being the domain. Therefore $\mathcal{C}_{k,q}$ is compact, if $k$ and $q$ are continuous and $\Omega$ is compact; this is because in this case the mapping $x \to q(x)k(\cdot, x)$ becomes continuous, and in general the image of a continuous mapping from a compact domain is compact. Thus, we make the following assumption:

Assumption 2. $\Omega$ is a compact metric space, $q : \Omega \to \mathbb{R}$ is continuous with $q(x) > 0$ for all $x \in \Omega$, and $k : \Omega \times \Omega \to \mathbb{R}$ is continuous.

The following simple lemma established a key connection between weak greedy algorithms and ABQ. For a geometric interpretation of (7) in terms of projections, see Fig. 1 in Appendix B.1

Lemma 3.1. (proof in Appendix B.1) Let $x_1, \ldots, x_n \in \Omega$ be such that the kernel matrix $K_n = (k(x_i, x_j))_{i,j=1}^{n} \in \mathbb{R}^{n \times n}$ is invertible. Define $h_x := q(x)k(\cdot, x)$ for any $x \in \mathcal{X}$, and let $S_n := \text{span}(h_{x_1}, \ldots, h_{x_n}) \subset \mathcal{H}_k$. Assume that $q(x) > 0$ holds for all $x \in \Omega$. Then for all $x \in \Omega$ we have
\[
q^2(x)k_{X_n}(x, x) = \text{dist}^2(h_x, S_n),
\]
where $k_{X_n}(x, x)$ is the GP posterior variance function given by (3). Moreover, we have
\[
e_n(\mathcal{C}_{k,q}) = \sup_{x \in \Omega} q(x)\sqrt{k_{X_n}(x, x)},
\]
where $e_n(\mathcal{C}_{k,q})$ is the worst case error defined by (5) with $\mathcal{C} := \mathcal{C}_{k,q}$ and $S_n$ defined here.

Lemma 3.1 (8) suggests that we can analyze the convergence properties of $\sup_{x \in \Omega} q(x)\sqrt{k_{X_n}(x, x)}$ for $X_n = \{x_1, \ldots, x_n\}$ generated from the ABQ rule (1) by analyzing those of the worst case error $e_n(\mathcal{C}_{k,q})$ for the corresponding elements $h_{x_1}, \ldots, h_{x_n}$, where $h_{x_i} := q(x_i)k(\cdot, x_i)$.
Adaptive Bayesian Quadrature as a Weak Greedy Algorithm. We now show that the ABQ [4] gives a weak greedy approximation of the compact set \( C_k \) in the RKHS \( H_k \) in the sense of [6]. We summarize required conditions in Assumptions [3] and [4] As mentioned above, Assumption [3] is the crucial one: its implications for certain specific ABQ methods will be discussed in Sec. 4.2.

Assumption 3 (Weak Adaptivity Condition). There are constants \( C_L, C_U > 0 \) such that \( C_L < b_n(x) < C_U \) holds for all \( x \in \Omega \) and for all \( n \in \mathbb{N} \cup \{0\} \).

Assumption 4. \( F : [0, \infty) \to [0, \infty) \) is increasing and continuous, and \( F(0) = 0 \). For any \( 0 < c \leq 1 \), there is a constant \( 0 < \psi(c) \leq 1 \) such that \( F^{-1} (cy) \geq \psi(c) F^{-1}(y) \) holds for all \( y \geq 0 \).

For instance, if \( F(y) = y^\delta \) for \( \delta > 0 \) then \( F^{-1}(y) = y^{1/\delta} \) and thus we have \( \psi(c) = c^{1/\delta} \) for \( 0 < c \leq 1 \); \( \delta = 1 \) is the case for the WSABI [13], and \( \delta > 0 \) for the VBMC [1, 2]. If \( F(y) = \exp(y) - 1 \) as in the MMLT [8], we have \( F^{-1}(y) = \log(y + 1) \) and it can be shown that \( \psi(c) = c \) for \( 0 < c \leq 1 \); see Appendix B.2. Note that in Assumption [4], the inverse \( F^{-1} \) is well-defined since \( F \) is increasing and continuous.

The following lemma guarantees we can assume without loss of generality that the kernel matrix \( K_n \) for the points \( x_1, \ldots, x_n \) generated from the ABQ [4] is invertible under the assumptions above, since otherwise \( \sup_{x \in \Omega} x_k(x, x) = 0 \) holds, implying that the quadrature error is 0 from Prop. 2.1. This guarantees the applicability of Lemma [3.1] for points generated from the ABQ [4].

Lemma 3.2. (proof in Appendix B.3) Suppose that Assumptions [2, 3] and [4] are satisfied. Then for \( x_1, \ldots, x_n \) generated from the ABQ [4], either one of the following holds: i) the kernel matrix \( K_\ell = (k(x_i, x_j))_{i,j=1}^\ell \in \mathbb{R}^{\ell \times \ell} \) is invertible for all \( \ell = 1, \ldots, n \); or ii) there exists some \( \ell = 1, \ldots, n \) such that \( \sup_{x \in \Omega} K_{\ell}(x, x) = 0 \).

Lemma [3.1] leads to the following theorem, which establishes a connection between ABQ and weak greedy algorithms.

Theorem 3.3. (proof in Appendix B.4) Suppose that Assumptions [2, 3] and [4] are satisfied. Let \( x_1, \ldots, x_n \) be generated by the ABQ [4], and define \( h_{x_i} = q(x_i)k(\cdot, x_i) \) for \( i = 1, \ldots, n \). Then \( h_{x_1}, \ldots, h_{x_n} \) are a \( \gamma \)-weak greedy approximation of \( C_k \) in \( H_k \) with \( \gamma = \sqrt{\psi(C_L/C_U)} \).

4 Convergence Rates of Adaptive Bayesian Quadrature

We use the connection established in the previous section to derive convergence rates of ABQ. To this end we introduce a quantity called Kolmogorov n-width, which is defined (for a Hilbert space \( H \) and a compact subset \( C \subset H \)) by

\[
d_n(C) := \inf_{U_n} \sup_{h \in C} \text{dist}(h, U_n),
\]

where the infimum is taken over all n-dimensional subspaces \( U_n \) of \( H \). This is the worst case error for the best possible solution using \( n \) elements in \( H \); thus \( d_n(C) \leq e_n(C) \) holds for any choice of \( S_n \) that defines the worst case error \( e_n(C) \) in [5]. The following result by DeVore et al. [11] Corollary 3.3] relates the Kolmogorov n-width with the worst case error \( e_n(C) \) of a weak greedy algorithm.

Lemma 4.1. Let \( H \) be a Hilbert space and \( C \subset H \) be a compact subset. For \( 0 < \gamma \leq 1 \), let \( h_1, \ldots, h_n \in C \) be a \( \gamma \)-weak greedy approximation of \( C \) in \( H \) for \( n \in \mathbb{N} \), and let \( e_n(C) \) be the worst case error [5] for the subspace \( S_n := \text{span}(h_1, \ldots, h_n) \). Then we have:

- Exponential decay: Assume that there exist constants \( \alpha > 0 \), \( C_0 > 0 \) and \( D_0 > 0 \) such that \( d_n(C) \leq C_0 \exp(-D_0 n^\alpha) \) holds for all \( n \in \mathbb{N} \). Then \( e_n(C) \leq \sqrt{2C_0} \gamma^{-1} \exp(-D_1 n^\alpha) \) holds for all \( n \in \mathbb{N} \) with \( D_1 := 2^{-1-\alpha} D_0 \).
– Polynomial decay: Assume that there exist constants $\alpha > 0$ and $C_0 > 0$ such that $d_n(C) \leq C_0 n^{-\alpha}$ holds for all $n \in \mathbb{N}$. Then $e_n(C) \leq C_1 n^{-\alpha}$ holds for all $n \in \mathbb{N}$ with $C_1 := 2^{5a+1+\gamma^2}C_0$.

Thus, the key is how to upper-bound the Kolmogorov $n$-width $d_n(C_{k,q})$ for the RKHS $H_k$ associated with the covariance kernel $k$. Given such an upper bound, one can then derive convergence rates for ABQ using Thm. 3.3. Below we demonstrate such results for kernels with infinite smoothness on $\Omega \subset \mathbb{R}^d$, such as Gaussian and (inverse) multiquadric kernels. In a similar way one can also derive rates for kernels with finite smoothness, such as Matérn and Wendland kernels. These additional results are presented in Appendix C.4.

4.1 Convergence Rates for Kernels with Infinite Smoothness

We consider kernels with infinite smoothness, such as square-exponential kernels $k(x,x') = \exp(-\|x-x'\|^2/\gamma^2)$ with $\gamma > 0$, multiquadric kernels $k(x,x') = (-1)^{[\beta]}(c^2 + \|x-x'\|^2)^{\beta}$ with $\beta, c > 0$ such that $[\beta]$ denotes the smallest integer greater than $\beta$, and inverse multiquadric kernels $k(x,x') = (c^2 + \|x-x'\|^2)^{-\beta}$ with $\beta > 0$. We have the following bound on the Kolmogorov $n$-width of the $C_{k,q}$ for these kernels; the proof is in Appendix C.2.

Proposition 4.2. Let $\Omega \subset \mathbb{R}^d$ be a cube, and suppose that Assumption 2 is satisfied. Let $k$ be a square-exponential kernel or an (inverse) multiquadric kernel. Then there exist constants $C_0, D_0 > 0$ such that $d_n(C_{k,q}) \leq C_0 \exp(-D_0 n^{1/d})$ holds for all $n \in \mathbb{N}$.

The requirement for $\Omega$ to be a cube stems from the use of Wendland [35, Thm. 11.22] in our proof, which requires this condition. In fact, this can be weakened to $O(\exp(-D_1 n^{-1/2d}))$ (note that this is still exponential); see [35, Sec. 11.4]. This also applies to the following results. Combining Prop. 4.2 with Lemma 3.1 Thm. 3.3 and Lemma 4.1, we now obtain a bound on $\sup_{x \in \Omega} \psi(x)$.

Theorem 4.3. (proof in Appendix C.3) Suppose that Assumptions 2, 3 and 4 are satisfied. Let $\Omega \subset \mathbb{R}^d$ be a cube, and $k$ be a square-exponential kernel or an (inverse) multiquadric kernel. Let $x_1, \ldots, x_n$ be generated by the ABQ [1]. Then there exist constants $C_1, D_1 > 0$ such that

$$\sup_{x \in \Omega} \psi(x) \sqrt{k_{x_n}(x,x)} \leq C_1 \psi(C_L/C_U)^{-1/2} \exp(-D_1 n^{1/d}) \quad (n \in \mathbb{N}).$$

As a directly corollary of Prop. 2.1 and Thm. 4.3, we finally obtain a convergence rate of the ABQ with an infinitely smooth kernel, which is exponentially fast.

Corollary 4.4. Suppose that Assumptions 1, 2, 3 and 4 are satisfied, and that $C_{\pi/q} := \int_{\Omega} |\pi(x)/q(x)|dx < \infty$. Let $\Omega \subset \mathbb{R}^d$ be a cube, and $k$ be a square-exponential kernel or an (inverse) multiquadric kernel. Let $X_n = \{x_1, \ldots, x_n\} \subset \Omega$ be generated by the ABQ [1]. Then there exists a constant $D_1 > 0$ independent of $n \in \mathbb{N}$ such that

$$\left| \int f(x) \pi(x) dx - \int T(m_{g,X_n}(x)) \pi(x) dx \right| = O(\exp(-D_1 n^{1/d})) \quad (n \to \infty).$$

4.2 Discussions of the Weak Adaptivity Condition (Assumption 3)

We discuss consequences of our results to individual ABQ methods reviewed in Sec. 2.2. We do this in particular by discussing the weak adaptivity condition (Assumption 3), which requires that the data-dependent term $b_n(x)$ in [1] is uniformly bounded away from zero and infinity. (A discussion for VBMC by Acerbi [10] 2
is given in Appendix [C.8]. To summarize, Assumption [3] holds if the densities of the variational distributions are bounded away uniformly from zero and infinity.

We first consider the WSABI-L approach by Gunter et al. [13], for which \( b_n(x) = (m_{g,X_n}(x))^2 \); a similar result is presented for the WSABI-M in Appendix C.7. The following bounds for \( b_n(x) \) follow from Lemma C.5 in Appendix C.5.

**Lemma 4.5.** Let \( b_n(x) = (m_{g,X_n}(x))^2 \). Suppose that Assumption [4] is satisfied, and that \( \inf_{x \in \Omega} |m(x)| > 2\|\tilde{g}\|_{H_k} k\|L_\infty(\Omega)\|^{1/2} \). Then Assumption [3] holds for \( C_L := \left( \inf_{x \in \Omega} |m(x)| - 2\|\tilde{g}\|_{H_k} k\|L_\infty(\Omega)\|^{1/2} \right)^2 > 0 \) and \( C_U := \left( \|m\|_{L_\infty(\Omega)} + 2\|\tilde{g}\|_{H_k} k\|L_\infty(\Omega)\|^{1/2} \right)^2 < \infty \).

Lemma 4.5 implies that WSABI-L may not be consistent when, e.g., one uses the zero prior mean function \( m(x) = 0 \), since in this case the condition \( \inf_{x \in \Omega} |m(x)| > 2\|\tilde{g}\|_{H_k} k\|L_\infty(\Omega)\|^{1/2} \) is not satisfied. Intuitively, the inconsistency may happen because the posterior mean \( m_{g,X_n}(x) \) for inputs \( x \) in regions distant from the current design points \( x_1, \ldots, x_n \) would become close to 0, since the prior mean function is 0; and such regions will never be explored in the subsequent iterations, because of the form \( b_n(x) = (m_{g,X_n}(x))^2 \). One simple way to guarantee the consistency is to make a modification like \( b_n(x) := \frac{1}{2}(m_{g,X_n}(x))^2 + \alpha = T(m_{g,X_n}(x)) \); then we can guarantee that \( C_L \geq \alpha > 0 \), encouraging exploration in the whole region \( \Omega \). This then makes the algorithm consistent.

We next consider the MMLT method by Chai and Garnett [8], for which \( b_n(x) = \exp(k_{X_n}(x) + 2m_{g,X_n}(x)) \). Lemma 4.6 below shows that the weak adaptivity condition holds for the MMLT as long as Assumption [4] is satisfied. Therefore different from the WSABI, the MMLT is consistent without requiring a further assumption.

**Lemma 4.6.** (proof in Appendix C.6) Let \( b_n(x) := \exp(k_{X_n}(x) + 2m_{g,X_n}(x)) \). Suppose that Assumption [4] is satisfied. Then Assumption [3] holds for \( C_L := \exp(-2\|m\|_{L_\infty(\Omega)} - 4\|\tilde{g}\|_{H_k} k\|L_\infty(\Omega)\|^{1/2}) > 0 \) and \( C_U := \exp(\|m\|_{L_\infty(\Omega)} + 2\|m\|_{L_\infty(\Omega)} + 4\|\tilde{g}\|_{H_k} k\|L_\infty(\Omega)\|^{1/2}) < 0 \).

## 5 Conclusion and Outlook

Extending efficient numerical integration beyond the low-dimensional domain remains both a formidable challenge and a crucial desideratum for many areas. In machine learning, efficient numerical integration in the high-dimensional domain would be a game-changer for Bayesian learning. Developed by, and used in, the machine learning community, adaptive Bayesian quadrature is a promising new direction for progress in this fundamental problem class. So far, it has been hindered by the absence of theoretical guarantees.

In this work, we have provided the first known convergence guarantees for ABQ methods, by analyzing a generic form of their acquisition functions. Of central importance is the notion of weak adaptivity which, speaking vaguely, ensures that the algorithm asymptotically does not “overly focus” on some evaluations. It is conceptually related to ideas like detailed balance and ergodicity, which play a similar role for Markov Chain Monte Carlo methods (where, speaking equally vaguely, they guard against the same kind of locality). Like those of MCMC, our sufficient conditions for consistency span a flexible class of design options, and can thus act as a guideline for the design of novel acquisition functions for ABQ, guided by practical and intuitive considerations. Based on the results presented herein, novel ABQ methods may be proposed for novel domains other than only positive integrands, for example integrands with discontinuities [23] and those with spatially inhomogeneous smoothness.

An important theoretical question, however, remains to be addressed: While our results provide convergence guarantees for ABQ methods, they do not provide a theoretical explanation for why, how and when ABQ methods should be fundamentally better than non-adaptive methods. In fact, little is known about theoretical
properties of adaptive quadrature methods in general. In applied mathematics, they remain an open problem [19,22]. While we have to leave this question of ABQ’s potential advantages over standard BQ for future research, we consider this area to be highly promising on account of the fundamental role of high-dimensional integrals of structured functions in probabilistic machine learning.

Acknowledgements

We thank Hans Kersting and Alexandra Gessner for fruitful discussions. This work has been supported by the ERC action StG 757275 / PANAMA.

A Appendices for Section 2

A.1 Proof of Prop. 2.1

In the proof we use the following notation: \( \|\sqrt{k}\|_{L_{\infty}} := \sup_{x \in \Omega} \sqrt{k(x,x)} \) and \( \|\sqrt{k_{X_n}}\|_{L_{\infty}} := \sup_{x \in \Omega} \sqrt{k_{X_n}(x,x)} \).

**Proof.** It is known that (see e.g. [15, Prop. 3.10]) the GP posterior standard deviation can be written as

\[
\sqrt{k_{X_n}(x,x)} = \sup_{u \in H_k : \|u\|_{H_k} \leq 1} |u(x) - k_n(x)^T K_n^{-1} u_n|, \quad x \in \Omega \tag{9}
\]

where \( u := (u(x_1), \ldots, u(x_n))^T \in \mathbb{R}^n \). Note that for any \( x \in \Omega \), we have \( m_{g,X_n}(x) = m(x) + k_n(x)^T K_n^{-1} \tilde{g}_n \), since \( \tilde{g}_n = (\tilde{g}(x_i))_{i=1}^n = (m(x_i) - g(x_i))_{i=1}^n = m_n - g_n \). Therefore by \( g(x) = m(x) + \tilde{g}(x) \), \( \tilde{g} \in H_k \) and \( \tilde{T} \) we have

\[
|g(x) - m_{g,X_n}(x)| = |\tilde{g}(x) - k_n(x)^T K_n^{-1} \tilde{g}| \leq \|\tilde{g}\|_{H_k} \sqrt{k_{X_n}(x,x)}. \tag{10}
\]

On the other hand, by Taylor’s theorem, there exists \( \alpha_{x,X_n} \in [0,1] \) such that for \( y_{x,X_n} := g(x) + \alpha_{x,X_n} (m_{g,X_n}(x) - g(x)) \in \mathbb{R} \) we have

\[
T(m_{g,X_n}(x)) = T(g(x)) + T'(y_{x,X_n}) (m_{g,X_n}(x) - g(x)),
\]

where \( T'(y) \) denotes the derivative of \( T \) at \( y \in \mathbb{R} \). From this and \( \|\tilde{T}\| \) we have

\[
|T(g(x)) - T(m_{g,X_n}(x))| \leq |T'(y_{x,X_n})| |m_{g,X_n}(x) - g(x)| \leq |T'(y_{x,X_n})| \|\tilde{g}\|_{H_k} \sqrt{k_{X_n}(x,x)}
\]

Note that \( |T'(y_{x,X_n})| \) is uniformly bounded over all \( x \in \Omega \) and \( n \in \mathbb{N} \), since \( T' \) is continuous by assumption and \( |y_{x,X_n}| \) is bound uniformly over all \( x \in \Omega \) and \( n \in \mathbb{N} \); the latter can be shown as

\[
|y_{x,X_n}| \leq |g(x)| + |\alpha_{x,X_n} (m_{g,X_n}(x) - g(x))| \leq |m(x)| + |\tilde{g}(x)| + |m_{g,X_n}(x) - g(x)| \\
\leq \|m\|_{L_{\infty}(\Omega)} + \|\tilde{g}\|_{H_k} \sqrt{k(x,x)} + \|\tilde{g}\|_{H_k} \sqrt{k_{X_n}(x,x)} \leq \|m\|_{L_{\infty}(\Omega)} + 2\|\tilde{g}\|_{H_k} \sqrt{k} \|_{L_{\infty}},
\]

where we used \( \|\tilde{g}\|_{H_k} = \|\tilde{g}, k(\cdot, x)\|_{H_k} \leq \|\tilde{g}\|_{H_k} \sqrt{k(x,x)} \) and \( k_{X_n}(x,x) \leq k(x,x) \). This implies that

\[
|T'(y_{x,X_n})| \leq \sup_{y \in \mathbb{R} : |y| \leq \|m\|_{L_{\infty}(\Omega)} + 2\|\tilde{g}\|_{H_k} \sqrt{k}} |T'(y)| =: C_{\tilde{g},m,k,T} < \infty.
\]

Therefore,

\[
|T(g(x)) - T(m_{g,X_n}(x))| \leq C_{\tilde{g},m,k,T} \|\tilde{g}\|_{H_k} \sqrt{k_{X_n}(x,x)},
\]
which implies that
\[ \left| \int T(g(x))\pi(x)dx - \int T(m_{g,X_n}(x))\pi(x)dx \right| \leq \int |T(g(x)) - T(m_{g,X_n}(x))|\pi(x)dx \]
\[ \leq C_{\tilde{g},m,k,T}\|\tilde{g}\|_{\mathcal{H}_k} \int \sqrt{k_{X_n}(x,x)}\pi(x)dx \]
\[ \leq C_{\tilde{g},m,k,T}C_{\pi,q}\|\tilde{g}\|_{\mathcal{H}_k} \sup_{x \in \Omega} q(x)\sqrt{k_{X_n}(x,x)}, \]
where the last inequality follows from Hölder’s inequality.

\[ \square \]

### B Appendices for Section 3

#### B.1 Proof of Lemma 3.1

**Proof.** It is easy to show by the reproducing property that the GP posterior variance \( k_{X_n}(x,x) \) in (8) can be written as the squared RKHS distance between \( k(\cdot,x) \) and its orthogonal projection onto \( \text{span}(k(\cdot,x_1),\ldots,k(\cdot,x_n)) \subset \mathcal{H}_k \), provided that the kernel matrix \( K_n = (k(x_i,x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n} \) is invertible:

\[ k_{X_n}(x,x) = \text{dist}^2(k(\cdot,x),\text{span}(k(\cdot,x_1),\ldots,k(\cdot,x_n))) = \inf_{\alpha_1,\ldots,\alpha_n \in \mathbb{R}} \|k(\cdot,x) - \sum_{i=1}^n \alpha_i k(\cdot,x_i)\|_{\mathcal{H}}^2. \]

Therefore,

\[ q^2(x)k_{X_n}(x,x) = \inf_{\alpha_1,\ldots,\alpha_n \in \mathbb{R}} \|q(x)k(\cdot,x) - \sum_{i=1}^n \alpha_i q(x)k(\cdot,x_i)\|_{\mathcal{H}}^2 \]
\[ = \inf_{\beta_1,\ldots,\beta_n \in \mathbb{R}} \|q(x)k(\cdot,x) - \sum_{i=1}^n \beta_i q(x_i)k(\cdot,x_i)\|_{\mathcal{H}}^2, \]
\[ = \inf_{g \in S_n} \|h_x - g\|_{\mathcal{H}}^2 = \text{dist}^2(h_x,S_n), \]

where the second equality follows from \( q(x) > 0 \) and \( q(x_i) > 0 \) for all \( i = 1,\ldots,n \); this proves (7). Using this, (5) and the definition of \( C_{k,q} \), the identity (8) can be shown as

\[ e_n(C_{k,q}) = \sup_{h \in C_{k,q}} \text{dist}(h_x,S_n) = \sup_{x \in \Omega} \text{dist}(h_x,S_n) = \sup_{x \in \Omega} q(x)\sqrt{k_{X_n}(x,x)}. \]

\[ \square \]

Fig. 1 provides a geometric interpretation of (7) in Lemma 3.1 and its proof.

#### B.2 An Example for Assumption 4

The following lemma gives the constant \( \psi(c) \) in Assumption 4 for the case \( F(y) = \exp(y) - 1 \), and thus \( F^{-1}(y) = \log(1+y) \), of the MMLT [8]: \( \psi(c) = c \). The proof is elementary, but we include it for completeness.

**Lemma B.1.** For any \( 0 < c \leq 1 \), we have \( \log(1+cy) \geq c\log(1+y) \) for all \( y \geq 0 \).

**Proof.** The assertion is equivalent to that \( 1+cy \geq (1+y)^c \) holds for all \( y \geq 0 \), which we show below. Let \( f(y) := 1+cy \) and \( g(y) := (1+y)^c \) for \( y \geq 0 \). Their derivatives are \( f'(y) = c \) and \( g'(y) = c(1+y)^{c-1} \), for which we have \( f'(y) \geq g'(y) \) for all \( y \geq 0 \), since \( c-1 \leq 0 \). We also have \( f(0) = g(0) = 1 \). Therefore, by the fundamental theorem of calculus, we conclude that \( f(y) = f(0) + \int_0^y f'(\tilde{y})d\tilde{y} \geq g(0) + \int_0^y g'(\tilde{y})d\tilde{y} = g(y) \) for all \( y \geq 0 \).

\[ \square \]
Assume that ii) does not hold. Then there exists \( y \) such that
\[ q(x)k(\cdot, x) = q(y)k(\cdot, y) \]
which implies
\[ q(\cdot) = q(y) \]
This expression and hence \( x = y \) follows from \( q(x_1), q(x_2) > 0 \).

**B.3 Proof of Lemma 3.2**

*Proof.* Let \( \ell = 1, \ldots, n - 1 \), and assume that \( x_1, \ldots, x_\ell \in \Omega \) are such that the kernel matrix \( K_\ell = (k(x_i, x_j))_{i,j=1}^{\ell} \in \mathbb{R}^{\ell \times \ell} \) is invertible; this is always true for \( \ell = 1 \). For \( x_{\ell+1} \in \arg \max_{x \in \Omega} a_\ell(x) = \arg \max_{x \in \Omega} F(q^2(x)k_{X_\ell}(x, x)) b_\ell(x) \), we show that either of the following holds: i) \( k(\cdot, x_{\ell+1}) \) is linearly independent to \( k(\cdot, x_1), \ldots, k(\cdot, x_\ell) \) and thus \( K_{\ell+1} = (k(x_i, x_j))_{i,j=1}^{\ell+1} \in \mathbb{R}^{(\ell+1) \times (\ell+1)} \) is invertible, or ii) \( \sup_{x \in \Omega} k_{X_\ell}(x, x) = 0 \).

Assume that ii) does not hold. Then there exists \( y \in \Omega \) such that \( k_{X_\ell}(y, y) > 0 \). For this \( y \) we have \( a_\ell(y) = F(q^2(y)k_{X_\ell}(x, x)) b_\ell(x) > 0 \), since \( q(x), b_\ell(x) > 0 \) for all \( x \in \Omega \), \( F(0) = 0 \) and \( F \) is increasing. Therefore \( a_\ell(x_{\ell+1}) \geq a_\ell(y) > 0 \), and thus \( k_{X_\ell}(x_{\ell+1}, x_{\ell+1}) > 0 \). Note that since the kernel matrix \( K_\ell \) is invertible, we have
\[ k_{X_\ell}(x_{\ell+1}, x_{\ell+1}) = \inf_{\alpha_1, \ldots, \alpha_\ell \in \mathbb{R}} \| k(\cdot, x_{\ell+1}) - \sum_{i=1}^\ell \alpha_i k(\cdot, x_i) \|^2_{K_\ell} \]
This expression and \( k_{X_\ell}(x_{\ell+1}, x_{\ell+1}) > 0 \) imply that \( k(\cdot, x_{\ell+1}) \) is linearly independent to \( k(\cdot, x_1), \ldots, k(\cdot, x_\ell) \), since otherwise \( k(\cdot, x_{\ell+1}) \) can be written as a linear combination of \( k(\cdot, x_1), \ldots, k(\cdot, x_\ell) \), and thus \( k_{X_\ell}(x_{\ell+1}, x_{\ell+1}) \) becomes 0 from the above expression. Thus i) has been shown.

**B.4 Proof of Theorem 3.3**

*Proof.* For \( \ell = 0, \ldots, n - 1 \), we have by \( x_{\ell+1} \in \arg \max_{x \in \Omega} a_\ell(x) \) that
\[ a_\ell(x_{\ell+1}) = F(q^2(x_{\ell+1})k_{X_\ell}(x_{\ell+1}, x_{\ell+1})) b_\ell(x_{\ell+1}) = \sup_{x \in \Omega} F(q^2(x)k_{X_\ell}(x, x)) b_\ell(x) \]
\[ \geq \sup_{x \in \Omega} F(q^2(x)k_{X_\ell}(x, x)) C_L = F \left( \sup_{x \in \Omega} q^2(x)k_{X_\ell}(x, x) \right) C_L, \]
where the last equality follows from $F$ being an increasing function. This implies that

$$F\left(q^2(x_{\ell+1})k_{X_\ell}(x_{\ell+1}, x_{\ell+1})\right) \geq (C_L/C_U)F\left(\sup_{x \in \Omega} q^2(x)k_{X_\ell}(x, x)\right)$$

and therefore, again by $F$ being increasing,

$$q^2(x_{\ell+1})k_{X_\ell}(x_{\ell+1}, x_{\ell+1}) \geq F^{-1}\left((C_L/C_U)F\left(\sup_{x \in \Omega} q^2(x)k_{X_\ell}(x, x)\right)\right)$$

$$\geq \psi(C_L/C_U) \sup_{x \in \Omega} q^2(x)k_{X_\ell}(x, x).$$

Note that $\|h_x\|_{\mathcal{H}_k}^2 = \|q(x)k(\cdot, x)\|_{\mathcal{H}_k}^2 = q^2(x)k(x, x)$ for all $x \in \Omega$. Therefore for $\ell = 0$, in which case $k_{X_0}(x, x) = k(x, x)$, we have $\|h_{x_1}\|_{\mathcal{H}_k} \geq \sqrt{\psi(C_L/C_U)} \sup_{x \in \Omega} \|h_x\|_{\mathcal{H}_k} = \sqrt{\psi(C_L/C_U)} \sup_{h \in \mathcal{C}_k} \|h\|_{\mathcal{H}_k}$. For $\ell = 1, \ldots, n-1$ we have by Lemma C.1

$$\text{dist}^2(h_{x_{\ell+1}}, S_\ell) = q^2(x_{\ell+1})k_{X_\ell}(x_{\ell+1}, x_{\ell+1})$$

$$\geq \psi(C_L/C_U) \sup_{h \in \mathcal{C}_k} \text{dist}^2(h, S_\ell) = \psi(C_L/C_U) \sup_{h \in \mathcal{C}_k} \text{dist}^2(h, S_\ell).$$

Thus (6) holds for $\gamma = \sqrt{\psi(C_L/C_U)}$, which completes the proof. \qed

C. Appendices for Section 4

C.1 A Bound on the Kolmogorov n-width

Lemma C.1. Let $x_1, \ldots, x_n \in \Omega$ be such that the kernel matrix $K_n = (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$ is invertible. Assume that $q(x) > 0$ for all $x \in \Omega$. Then we have $d_n(C_{k,q}) \leq \inf_{x_1, \ldots, x_n \in \Omega} \sup_{x \in \Omega} q(x) \sqrt{k_{X_n}(x, x)}$.

Proof. Using Lemma 3.1 the Kolmogorov n-width can be upper-bounded as

$$d_n(C_{k,q}) = \inf_{U_n} \sup_{h \in \mathcal{C}_k} \text{dist}(h, U_n) = \inf_{U_n} \sup_{h \in \mathcal{C}_k} \text{dist}(h_x, U_n)$$

$$\leq \inf_{x_1, \ldots, x_n \in \Omega} \sup_{x \in \Omega} \text{dist}(h_x, S_n) = \inf_{x_1, \ldots, x_n \in \Omega} \sup_{x \in \Omega} q(x) \sqrt{k_{X_n}(x, x)},$$

where the infimum in the first line is taken over all $n$-dimensional subspaces $U_n$ of $\mathcal{H}_k$, and $S_n = \text{span}(h_{x_1}, \ldots, h_{x_n})$ with $h_x = q(x)k(\cdot, x)$. \qed

Lemma C.1 can be used for deriving upper-bounds on the Kolmogorov n-width $d_n(C_{k,q})$ for concrete examples of the kernel $k$ on $\Omega \subset \mathbb{R}^d$. To this end, the key quantity is the fill distance defined by

$$h_{X_n,\Omega} := \sup_{x \in \Omega} \min_{i=1, \ldots, n} \|x - x_i\|,$$

where $X_n := \{x_1, \ldots, x_n\} \subset \Omega$. This measures how densely the points $x_1, \ldots, x_n$ fill the region $\Omega$.

C.2 Proof of Prop. 4.2 (Kolmogorov n-width for kernels with infinite smoothness)

Proof. By [35, Theorem 11.22], where $k_{X_n}(x, x)$ is called the power function, there is a constant $c > 0$ such that $k_{X_n}(x, x) \leq \exp(-c_1/h_{X_n,\Omega})$ holds for any set of design points $X_n = \{x_1, \ldots, x_n\}$ with sufficiently small
If we define \( x_1, \ldots, x_n \) as equally-spaced grid points in \( \Omega \), then we have \( h_{X_n, \Omega} = C_2 n^{-1/d} \) for some \( C_2 > 0 \) independent of \( n \). Therefore for large enough \( n \), we have \( k_{X_n}(x, x') \leq \exp(-c_1 C_2^{-1} n^{1/d}) \). Note that there exists a constant \( c_3 > 0 \) such that \( k_{X_n}(x, x') \leq c_3 \) holds for all \( x \in \Omega \) and for all \( n \), since \( \Omega \) is compact and \( k_{X_n}(x, x) \) is continuous w.r.t. \( x \) for any fixed \( n \) and non-increasing w.r.t. \( n \) for any fixed \( x \in \Omega \). Therefore we have \( k_{X_n}(x, x) \leq c_3 \exp(-c_1 C_2^{-1} n^{1/d}) \) for all \( x \in \Omega \) and for all \( n \in \mathbb{N} \), if \( x_1, \ldots, x_n \) are equally-spaced grid points.

Note that \( \inf_{x_1, \ldots, x_n \in \Omega} \sup_{x \in \Omega} q(x) \sqrt{k_{X_n}(x, x)} \leq \sup_{x \in \Omega} q(x) \sqrt{k_{X_n}(x, x)} \) holds for any fixed choice of \( x_1, \ldots, x_n \) defining \( k_{X_n}(x, x) \) in the upper-bound. If we chose \( x_1, \ldots, x_n \) as equally-spaced grid points in the upper-bound, we have by Lemma C.1 and the above argument that

\[
d_n(C_{k,q}) \leq \inf_{x_1, \ldots, x_n \in \Omega} \sup_{x \in \Omega} q(x) \sqrt{k_{X_n}(x, x)} \leq \sup_{x \in \Omega} q(x) \sqrt{k_{X_n}(x, x)} \leq \sup_{x \in \Omega} q(x) \exp(-\frac{1}{2} c_1 C_2^{-1} n^{1/d}).
\]

Setting \( C_0 := \sup_{x \in \Omega} q(x) \sqrt{c_3} \) and \( D_0 := \frac{1}{2} c_1 C_2^{-1} \) concludes the proof.

C.3 Proof of Theorem 4.3

Proof. By Thm. 3.3, \( h_{x_1}, \ldots, h_{x_n} \) are \( \gamma \)-weak approximations of \( C_{k,q} \) in \( H_k \) with \( \gamma = \sqrt{\psi(C_L/C_U)} \). From this, and by Lemma 4.1 (exponential) and Prop. 4.2, there exist \( C_0, D_0 > 0 \) such that for \( C_1 := 2C_0 \) and \( D_1 := 2^{-1/2} D_0 \), we have \( e_n(C_{k,q}) \leq C_1 \psi(C_L/C_U) \exp(-D_1 n^{-1/d}) \) for all \( n \in \mathbb{N} \). Combining this and \( \psi \) in Lemma 3.1 concludes the proof.

C.4 Convergence Rates for ABQ using Kernels with Finite Smoothness

We deal with here kernels with finite smoothness. In particular, we consider shift-invariant kernels of the form

\[
k(x, x') = \Phi(x - x') \quad \text{with} \quad \Phi \in L_1(\mathbb{R}^d)
\]

satisfying

\[
c_1 (1 + \|\omega\|^2)^{-r} \leq \tilde{\Phi}(\omega) \leq c_2 (1 + \|\omega\|^2)^{-r}, \quad \omega \in \mathbb{R}^d
\]

for some \( c_1, c_2 > 0 \) and \( r > d/2 \), where \( \tilde{\Phi} \) denotes the Fourier transform of \( \Phi \). The RKHS of such a kernel is norm-equivalent to a Sobolev space of order \( r \), which consists of functions whose weak derivative up to order \( r \) exist and are square-integrable [35 Corollary 10.48]; thus \( r \) represents the smoothness of functions in the RKHS.

For instance, Matérn kernels [29, p. 84] of the form

\[
k(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu}\|x - x'\|}{\ell} \right)^\nu K_\nu \left( \frac{\sqrt{2\nu}\|x - x'\|}{\ell} \right), \quad (\nu, \ell > 0)
\]

where \( \Gamma \) is the Gamma function and \( K_\nu \) is the modified Bessel function of third kind, satisfy \((11)\) with \( r = \nu + d/2 \). Another example is Wendland kernels [35 Theorem 10.35], which have compact supports and thus have computational advantages; see [33] and [35 Chapter 9] for details. In the following result, we use the notion of a Lipschitz boundary and an interior cone condition, the definitions of which can be found in, e.g., [17] Section 3 and references therein.

Assumption 5. \( \Omega \subset \mathbb{R}^n \) is a compact set having a Lipschitz boundary and satisfying an interior cone condition.

C.4.1 Kolmogorov n-width for kernels with finite smoothness

Proposition C.2. Suppose that Assumptions 3 and 5 are satisfied. Let \( k(x, x') = \Phi(x - x') \) be a kernel satisfying \((11)\) for \( r > d/2 \). Then there exists a constant \( C_0 > 0 \) such that

\[
d_n(C_{k,q}) \leq C_0 n^{-r/d+1/2}, \quad n \in \mathbb{N}.
\]
exists a constant \(ABQ\) \((4)\).

**Theorem C.3.** Suppose that Assumptions 2, 3, 4 and 5 are satisfied. Let 

\[
\text{as a corollary of Prop. 2.1 and Thm. C.3, we have the following result.}
\]

**Proof.** Let \(\sup_{x \in \Omega} \Phi'(x, x') \leq \phi(x, x') \leq \phi(x, x') \leq \phi(x, x')\). By setting \(x, x, x, x\) as equally-spaced grid points in \(\Omega\), there exists a constant \(c_2 > 0\) such that \(h_{x, x} \leq c_2 n^{-1/d}\). Therefore we have for some \(c_3 > 0\)

\[
\sup_{g \in H_k; \|g\|_{H_k} \leq 1} \|g - m_{g, X_n}\|_{L_\infty(\Omega)} \leq c_3 n^{-r/d+1/2}
\]

for sufficiently large \(n\). Note that the GP posterior variance can be written as (see e.g. \[15, \text{Prop. 3.10}\])

\[
\sqrt{k_{X_n}(x, x)} = \sup_{g \in H_k; \|g\|_{H_k} \leq 1} |g(x) - m_{g, X_n}(x)|, \quad x \in \Omega.
\]

This implies that \(\sqrt{k_{X_n}(x, x)} \leq \sup_{\|g\|_{H_k} \leq 1} \|g - m_{g, X_n}\|_{L_\infty(\Omega)}\) for all \(x \in \Omega\), which further implies that \(\sup_{x \in \Omega} \sqrt{k_{X_n}(x, x)} \leq \sup_{\|g\|_{H_k} \leq 1} \|g - m_{g, X_n}\|_{L_\infty(\Omega)}\). Therefore, for large enough \(n\) we have \(\sup_{x \in \Omega} \sqrt{k_{X_n}(x, x)} \leq c_3 n^{-r/d+1/2}\) if \(x, x, x, x\) are equally-spaced grid points in \(\Omega\). Note that there exists a constant \(c_4 > 0\) such that \(k_{X_n}(x, x) \leq c_4\) holds for all \(x \in \Omega\) and for all \(n \in \mathbb{N}\), since \(\Omega\) is compact, \(k_{X_n}(x, x)\) is continuous w.r.t. \(x\) for any fixed \(n\) and \(k_{X_n}(x, x)\) is non-increasing w.r.t. \(n\) for any fixed \(x \in \Omega\). By setting \(c_5 := \max(c_3, c_4)\), this implies that, for every \(n\), we have

\[
\sup_{x \in \Omega} \sqrt{k_{X_n}(x, x)} \leq c_5 n^{-r/d+1/2}\]

if \(x, x, x, x\) are equally-spaced grid points in \(\Omega\). Finally, by \(\text{Lemma C.1}\) we have

\[
d_n(C_{k, q}) \leq \sup_{x \in \Omega} q(x) c_5 n^{-r/d+1/2}
\]

and thus the assertion holds with \(C_0 := \sup_{x \in \Omega} q(x) c_5 < \infty\), which is bounded since \(q\) is continuous and \(\Omega\) is compact.  

\(\square\)

**C.4.2 Convergence Rates**

Combining Prop. C.2 and Thm. 3.3, we have the following bound on \(\sup_{x \in \Omega} q(x) \sqrt{k_{X_n}(x, x)}\), when \(x, x, x, x\) are generated by the ABQ \[4\].

**Theorem C.3.** Suppose that Assumptions 2, 3, 4 and 5 are satisfied. Let \(k(x, x') = \Phi(x - x')\) be a kernel satisfying \((11)\) for \(r > d/2\). For \(n \in \mathbb{N}\), let \(X_n = \{x_1, \ldots, x_n\} \subset \Omega\) be generated by the ABQ \[4\]. Then there exists a constant \(C_1 > 0\) such that

\[
\sup_{x \in \Omega} q(x) \sqrt{k_{X_n}(x, x)} \leq C_1 \psi(C_L/C_U)^{-1} n^{-r/d+1/2}, \quad n \in \mathbb{N}.
\]

**Proof.** Let \(h_x := q(x) k(. . x)\) for any \(x \in \Omega\). Then by Thm. 3.3 \(h_x, \ldots, h_{x_n}\) are a \(\gamma\)-weak greedy approximation of \(C_{k, q}\) in \(H_k\) with \(\gamma = \sqrt{\psi(C_L/C_U)}\). From this, and by \(\text{Lemma 4.1}\) (polynomial decay) and Prop. C.2 there exists a constant \(C_0 > 0\) such that \(e_n(C_{k, q}) \leq 2^{5a+2}\gamma^{-2} C_0 n^{-\alpha}\) holds for all \(n \in \mathbb{N}\), where \(\alpha := r/d - 1/2\). Combining this inequality and \[8\] yields assertion with \(C_1 = 2^{5a+2} C_0\).  

\(\square\)

As a corollary of Prop. 2.1 and Thm. C.3, we have the following result.

**Corollary C.4.** Suppose that Assumptions 2, 3, 4 and 5 are satisfied, and that \(C_{n/q} := \int |\pi(x)/q(x)| dx < \infty\). Assume \(k(x, x') = \Phi(x - x')\) satisfies \((11)\) with \(r > d/2\). For \(n \in \mathbb{N}\), let \(x_1, \ldots, x_n\) be generated by the ABQ \[4\]. Then we have

\[
\left| \int f(x) \pi(x) dx - \int T(m_{g, X_n}(x)) \pi(x) dx \right| = O(n^{-r/d+1/2}) \quad (n \to \infty)
\]
C.5 Bounds for GP Posterior Mean Functions

The following lemma is used for deriving the constants \(C_L\) and \(C_U\) in Assumption 3 for individual ABQ methods.

**Lemma C.5.** Assume that \(\tilde{g} := g - m \in \mathcal{H}_k\). Then for all \(x \in \Omega\) and \(n \in \mathbb{N}\), we have

\[
|m(x)| - 2\|\tilde{g}\|_{\mathcal{H}_k} \sqrt{k(x,x)} \leq |m_{g,X_n}(x)| \leq |m(x)| + 2\|\tilde{g}\|_{\mathcal{H}_k} \sqrt{k(x,x)}
\]

**Proof.** We show the lower-bound; the upper-bound can be shown similarly. Since \(\tilde{g} \in \mathcal{H}_k\), we have

\[
|m_{\tilde{g},X_n}(x)| \leq |\tilde{g}(x)| + |\tilde{g}(x) - m_{\tilde{g},X_n}(x)|
\]

\[
\leq \|\tilde{g}\|_{\mathcal{H}_k} \sqrt{k(x,x)} + \|\tilde{g}\|_{\mathcal{H}_k} \sqrt{k_{X_n}(x,x)} \leq 2\|\tilde{g}\|_{\mathcal{H}_k} \sqrt{k(x,x)}.
\]

Note that \(m_{g,X_n}(x) = m(x) + m_{\tilde{g},X_n}(x)\) since \(\tilde{g} = g - m\). Therefore,

\[
|m_{g,X_n}(x)| \geq |m(x)| - |m_{\tilde{g},X_n}(x)| \geq |m(x)| - 2\|\tilde{g}\|_{\mathcal{H}_k} \sqrt{k(x,x)}.
\]

\[\square\]

C.6 Proof of Lemma 4.6

**Proof.** First note that \(0 \leq k_{X_n}(x,x) \leq k(x,x)\) for all \(x \in \Omega\) and \(n \in \mathbb{N}\). Using Lemma C.5, we have

\[
\exp(k_{X_n}(x,x) + 2m_{g,X_n}(x)) \leq \exp(k_{X_n}(x,x) + 2|m_{g,X_n}(x)|)
\]

\[
\leq \exp(k(x,x) + 2|m(x)| + 4\|\tilde{g}\|_{\mathcal{H}_k} \sqrt{k(x,x)})
\]

\[
\leq \exp(\|k\|_{L_\infty(\Omega)} + 2\|m\|_{L_\infty(\Omega)} + 4\|\tilde{g}\|_{\mathcal{H}_k} \|k\|_{L_\infty(\Omega)}^{1/2})
\]

Similarly, we have

\[
\exp(k_{X_n}(x,x) + 2m_{g,X_n}(x)) \geq \exp(2m_{g,X_n}(x))
\]

\[
\geq \exp(-2|m_{X_n}(x)|)
\]

\[
\geq \exp(-2|m(x)| - 4\|\tilde{g}\|_{\mathcal{H}_k} \sqrt{k(x,x)})
\]

\[
\geq \exp(-2\|m\|_{L_\infty(\Omega)} - 4\|\tilde{g}\|_{\mathcal{H}_k} \|k\|_{L_\infty(\Omega)}^{1/2})
\]

\[\square\]

C.7 Bounds for WSABI-M

The following bounds for \(b_{X_n}(x) = \frac{1}{2}k_{X_n}(x,x) + (m_{g,X}(x))^2\) of the WSABI-M [13] can be easily obtained using Lemma C.5, and \(0 \leq \frac{1}{2}k_{X_n}(x,x) \leq \frac{1}{2}\|k\|_{L_\infty(\Omega)}\).

**Lemma C.6.** Let \(b_{X_n}(x) = \frac{1}{2}k_{X_n}(x,x) + (m_{g,X}(x))^2\). Suppose that Assumption 7 is satisfied, and that

\[
\inf_{x \in \Omega} |m(x)| > 2\|\tilde{g}\|_{\mathcal{H}_k} \|k\|_{L_\infty(\Omega)}^{1/2}
\]

Then \(C_L < b_n(x) < C_U\) for all \(x \in \Omega\) and \(n \in \mathbb{N}\), where \(C_L := (\inf_{x \in \Omega} |m(x)| - 2\|\tilde{g}\|_{\mathcal{H}_k} \|k\|_{L_\infty(\Omega)}^{1/2})^2 > 0\) and \(C_U := \frac{1}{2}\|k\|_{L_\infty(\Omega)} + (\|m\|_{L_\infty(\Omega)} + \|k\|_{L_\infty(\Omega)}^{1/2})^2 < \infty\).
C.8 Discussion for Variational Bayesian Monte Carlo (VBMC)

The VBMC by Acerbi [1, 2] uses \( F(y) = y^3 \), \( q(x) = 1 \) and \( b_n(x) = \pi_n^{\delta_2}(x) \exp(\delta_3 m_g, x_n(x)) \), where \( \pi_n \) is the variational posterior at the \( n \)-th iteration and \( \delta_1, \delta_2, \delta_3 \geq 0 \) are constants. Recall that in this method the transformation is identity: \( T(y) = y \) for \( y \in \mathbb{R} \); thus \( g = f \). The following result can be easily obtained from Lemma C.5.

**Lemma C.7.** Let \( b_n(x) = \pi_n^{\delta_2}(x) \exp(\delta_3 m_g, x_n(x)) \) with \( \delta_2, \delta_3 \geq 0 \). Suppose that Assumption 1 is satisfied, and that there exist constants \( D_L, D_U \) such that \( 0 < D_L < \pi_n(x) < D_U < \infty \) holds for all \( x \in \Omega \) and \( n \in \mathbb{N} \). Then we have \( C_L < b_n(x) < C_U < \infty \) for all \( x \in \Omega \) and \( n \in \mathbb{N} \), where \( C_L := D_L^{\delta_2} \exp(-\delta_3(\|m\|_{L^\infty(\Omega)} + 2\|\tilde{g}\|_{H_k} \|k\|_{L^\infty(\Omega)}^{1/2})) \geq 0 \) and \( C_U := D_U^{\delta_2} \exp(\delta_3(\|m\|_{L^\infty(\Omega)} + 2\|\tilde{g}\|_{H_k} \|k\|_{L^\infty(\Omega)}^{1/2})) < \infty \).

The condition \( 0 < D_L < \pi_n(x) < D_U < \infty \) for all \( x \in \Omega \) and \( n \in \mathbb{N} \) requires that i) the supports of the variational distributions should cover the whole domain \( \Omega \); and that ii) the density values of the variational distributions should be uniformly bounded from above. This implies that, if the variational family is a set of Gaussian mixtures (as proposed by Acerbi [1, 2]), then the variance of each mixture component should be uniformly lower- and upper-bounded; otherwise the condition \( 0 < D_L < \pi_n(x) < D_U < \infty \) may not be satisfied.

We note that in the setting of the VBMC, the density \( \pi \) in the target integral \( \int f(x) \pi(x) dx \) is an intractable posterior density, and it is to be approximated as \( \int m_f, x_n(x) \pi_n(x) dx \) using the variational posterior \( \pi_n \); therefore there is also an error due to the approximation of \( \pi \) by \( \pi_n \). Thus, a complete theoretical analysis requires analyzing the convergence behavior of the variational posterior \( \pi_n \); this is out of scope of this paper and we leave it for future research.

References


