Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings

Motonobu Kanagawa

Joint work with
Bharath Sriperubudur (Penn State) and Kenji Fukumizu (ISM)

SAMSI Reading Group on Probabilistic Numerics,
16 October 2017

Outline

Outline

Numerical integration (quadrature)

A fundamental task in various fields, including machine learning and statistics.

P: a known probability measure on $\Omega \subset \mathbb{R}^d$.

f: an integrand on Ω.

$\int f(x)dP(x)$: the integral, assumed having no closed form solution.

Numerical integration (quadrature)

A fundamental task in various fields, including machine learning and statistics.

P: a known probability measure on $\Omega \subset \mathbb{R}^d$.

f: an integrand on Ω.

$\int f(x)dP(x)$: the integral, assumed having no closed form solution.

- The task is to approximate $\int f(x)dP(x)$ in the form

$$\sum_{i=1}^{n} w_i f(X_i) \approx \int f(x)dP(x),$$

- How to construct weighted points $\{(w_i, X_i)\}_{i=1}^n$?

Numerical integration (quadrature)

A fundamental task in various fields, including machine learning and statistics.

P: a known probability measure on $\Omega \subset \mathbb{R}^d$.

f: an integrand on Ω.

$\int f(x)dP(x)$: the integral, assumed having no closed form solution.

- The task is to approximate $\int f(x)dP(x)$ in the form

$$\sum_{i=1}^{n} w_i f(X_i) \approx \int f(x)dP(x),$$

- How to construct weighted points $\{(w_i, X_i)\}_{i=1}^n$?
- e.g. Monte Carlo generates $X_1, \ldots, X_n$ randomly, and uses importance weights as $w_1, \ldots, w_n$.

  $\Rightarrow$ slow convergence rate $O(n^{-1/2})$:(

Notation throughout the presentation

Please remember this!

$$Pf := \int f(x) dP(x), \quad P_n f := \sum_{i=1}^{n} w_i f(X_i).$$

Kernel-based quadrature rules

Let k be a kernel on Ω, and $\mathcal{H}_k$ be its reproducing kernel Hilbert space (RKHS).

Kernel-based quadrature rules

Let k be a kernel on Ω, and $\mathcal{H}_k$ be its reproducing kernel Hilbert space (RKHS).

Worst case error

- In kernel quadrature, $\{(w_i, X_i)\}_{i=1}^n$ is constructed so that the worst case error in the RKHS

$$e_n(P; \mathcal{H}_k) := \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |Pf - P_n f|$$

  is to minimized.

Kernel-based quadrature rules

Let k be a kernel on Ω, and $\mathcal{H}_k$ be its reproducing kernel Hilbert space (RKHS).

Worst case error

- In kernel quadrature, $\{(w_i, X_i)\}_{i=1}^n$ is constructed so that the worst case error in the RKHS

$$e_n(P; \mathcal{H}_k) := \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |Pf - P_n f|$$

  is to minimized.

Examples of kernel quadrature

- Quasi Monte Carlo [Hickernell, 1998, Dick et al., 2013]
- Bayesian quadrature [O'Hagan, 1991, Briol et al., 2016]
- Kernel herding [Chen et al., 2010, Bach et al., 2012]

Kernel-based quadrature rules

Consider a well-specified case where $f \in \mathcal{H}_k$.

Kernel-based quadrature rules

Consider a well-specified case where $f \in \mathcal{H}_k$.

- The quadrature error is bounded as

$$|P_n f - Pf| \leq \underbrace{\|f\|_{\mathcal{H}_k}}_{RKHS\ \ norm} \times \underbrace{e_n(P; \mathcal{H}_k)}_{Worst\ \ case\ \ error} .$$

Kernel-based quadrature rules

Consider a well-specified case where $f \in \mathcal{H}_k$.

- The quadrature error is bounded as

$$|P_n f - Pf| \leq \underbrace{\|f\|_{\mathcal{H}_k}}_{RKHS\ \ norm} \times \underbrace{e_n(P; \mathcal{H}_k)}_{Worst\ \ case\ \ error}.$$

- If $e_n(P; \mathcal{H}_k) = O(n^{-b})$ with $b > 0$, then

$$|P_n f - Pf| = O(n^{-b}) \quad (n \to \infty).$$

Kernel-based quadrature rules

Consider a well-specified case where $f \in \mathcal{H}_k$.

- The quadrature error is bounded as

$$|P_n f - Pf| \leq \underbrace{\|f\|_{\mathcal{H}_k}}_{RKHS\ \ norm} \times \underbrace{e_n(P; \mathcal{H}_k)}_{Worst\ \ case\ \ error}.$$

- If $e_n(P; \mathcal{H}_k) = O(n^{-b})$ with $b > 0$, then

$$|P_n f - Pf| = O(n^{-b}) \quad (n \to \infty).$$

- e.g. if $\mathcal{H}_k$ is the r-th order Sobolev space on $\mathbb{R}^d$, the optimal quadrature achieves $b = r/d$.
- If $r > d/2$, we have $b > 1/2$, so the rate is faster than Monte Carlo.

Kernel-based quadrature rules

Consider a well-specified case where $f \in \mathcal{H}_k$.

- The quadrature error is bounded as

$$|P_n f - Pf| \leq \underbrace{\|f\|_{\mathcal{H}_k}}_{RKHS\ \ norm} \times \underbrace{e_n(P; \mathcal{H}_k)}_{Worst\ \ case\ \ error}.$$

- If $e_n(P; \mathcal{H}_k) = O(n^{-b})$ with $b > 0$, then

$$|P_n f - Pf| = O(n^{-b}) \quad (n \to \infty).$$

- e.g. if $\mathcal{H}_k$ is the r-th order Sobolev space on $\mathbb{R}^d$, the optimal quadrature achieves $b = r/d$.
- If $r > d/2$, we have $b > 1/2$, so the rate is faster than Monte Carlo.
- Assumption $f \in \mathcal{H}_k$ reflects the knowledge of f being smooth.

Our focus: misspecified settings

Consider a misspecified setting where $f \notin \mathcal{H}_k$.

- Theoretical guarantees are no longer provided for kernel quadrature

Our focus: misspecified settings

Consider a misspecified setting where $f \notin \mathcal{H}_k$.

- Theoretical guarantees are no longer provided for kernel quadrature

When it occurs

- When one only has limited knowledge about f.
- e.g. when f is a black box function.

Our focus: misspecified settings

Consider a misspecified setting where $f \notin \mathcal{H}_k$.

- Theoretical guarantees are no longer provided for kernel quadrature

When it occurs

- When one only has limited knowledge about f.
- e.g. when f is a black box function.
  - The mapping $x \to f(x)$ may involve complicated simulation or real world experiments.
  - e.g. applications to computer graphics [Briol et al., 2016], marginal likelihood computation [Oates et al., 2016b].

Our focus: misspecified settings

Consider a misspecified setting where $f \notin \mathcal{H}_k$.

- Theoretical guarantees are no longer provided for kernel quadrature

When it occurs

- When one only has limited knowledge about f.
- e.g. when f is a black box function.
  - The mapping $x \to f(x)$ may involve complicated simulation or real world experiments.
  - e.g. applications to computer graphics [Briol et al., 2016], marginal likelihood computation [Oates et al., 2016b].

Questions

- Can we still guarantee the convergence of kernel quadrature?
- If so, do we need any condition on $\{(w_i, X_i)\}_{i=1}^n$?

Contributions

Settings

- Generic convergence analysis of deterministic kernel quadrature in misspecified settings
- Focus on Sobolev spaces as RKHSs.
  - Standard RKHSs, with kernels being Matérn's or Wendland's.

Contributions

Settings

- Generic convergence analysis of deterministic kernel quadrature in misspecified settings
- Focus on Sobolev spaces as RKHSs.
  - Standard RKHSs, with kernels being Matérn's or Wendland's.

Specific contributions

- Provide two different conditions for $\{(w_i, X_i)\}_{i=1}^n$ to derivate convergence rates.
- Convergence rates for Bayesian quadrature in misspecified settings.

Outline

Reproducing kernel Hilbert spaces (RKHS) [Aronszajn, 1950]

Let Ω be a set, and $k : \Omega \times \Omega \to \mathbb{R}$ be a positive definite kernel.

Reproducing kernel Hilbert spaces (RKHS) [Aronszajn, 1950]

Let Ω be a set, and $k : \Omega \times \Omega \to \mathbb{R}$ be a positive definite kernel.

- For k, there is a uniquely associated RKHS $\mathcal{H}_k$.
- $\mathcal{H}_k$ is a Hilbert space of functions on Ω, and satisfies
  1. For all $x \in \Omega$,
  $$k(\cdot, x) \in \mathcal{H}_k$$
  2. For all $f \in \mathcal{H}_k$ and $x \in \Omega$,
  $$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}$$

Reproducing kernel Hilbert spaces (RKHS) [Aronszajn, 1950]

Let Ω be a set, and $k : \Omega \times \Omega \to \mathbb{R}$ be a positive definite kernel.

- For k, there is a uniquely associated RKHS $\mathcal{H}_k$.
- $\mathcal{H}_k$ is a Hilbert space of functions on Ω, and satisfies
  1. For all $x \in \Omega$,
  $$k(\cdot, x) \in \mathcal{H}_k$$
  2. For all $f \in \mathcal{H}_k$ and $x \in \Omega$,
  $$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}$$
- If $k(x, y) = \Phi(x - y)$ for some $\Phi : \mathbb{R}^d \to \mathbb{R}$, the RKHS (on $\mathbb{R}^d$) is given by

$$\mathcal{H}_k = \left\{ f \in L_2(\mathbb{R}^d) : \|f\|_{\mathcal{H}_k}^2 := \int |\hat{f}(\xi)|^2 \hat{\Phi}(\xi)^{-1} d\xi < \infty \right\},$$

where $\hat{f}$ and $\hat{\Phi}$ are the Fourier transforms of f and Φ, respectively.

Sobolev spaces [Adams and Fournier, 2003]

A Sobolev space of order $r > d/2$ on $\mathbb{R}^d$ is defined by

$$H^r(\mathbb{R}^d) := \left\{ f \in L_2(\mathbb{R}^d) : \int |\hat{f}(\xi)|^2 (1 + \|\xi\|^2)^r \, d\xi < \infty \right\}.$$

Sobolev spaces [Adams and Fournier, 2003]

A Sobolev space of order $r > d/2$ on $\mathbb{R}^d$ is defined by

$$H^r(\mathbb{R}^d) := \left\{ f \in L_2(\mathbb{R}^d) : \int |\hat{f}(\xi)|^2 (1 + \|\xi\|^2)^r d\xi < \infty \right\}.$$

- $H^r(\mathbb{R}^d)$ is a Hilbert space with the inner-product

$$\langle f, g \rangle_{H^r(\mathbb{R}^d)} := \int \hat{f}(\xi)\overline{\hat{g}(\xi)}(1 + \|\xi\|^2)^r d\xi, \quad f, g \in H^r(\mathbb{R}^d).$$

Sobolev spaces [Adams and Fournier, 2003]

A Sobolev space of order $r > d/2$ on $\mathbb{R}^d$ is defined by

$$H^r(\mathbb{R}^d) := \left\{ f \in L_2(\mathbb{R}^d) : \int |\hat{f}(\xi)|^2 (1 + \|\xi\|^2)^r d\xi < \infty \right\}.$$

- $H^r(\mathbb{R}^d)$ is a Hilbert space with the inner-product

$$\langle f, g \rangle_{H^r(\mathbb{R}^d)} := \int \hat{f}(\xi)\overline{\hat{g}(\xi)}(1 + \|\xi\|^2)^r d\xi, \quad f, g \in H^r(\mathbb{R}^d).$$

- The order r quantifies the smoothness of functions in $H^r(\mathbb{R}^d)$.
- Each $f \in H^r(\mathbb{R}^d)$ has square-integrable (weak) derivatives up to order r.

Sobolev spaces [Adams and Fournier, 2003]

A Sobolev space $H^r(\Omega)$ on a measurable set $\Omega \subset \mathbb{R}^d$ is defined by the restriction of $H^r(\mathbb{R}^d)$;

$$H^r(\Omega) := \left\{ f : \Omega \to \mathbb{R} : f = g|_\Omega,\ \exists\, g \in H^r(\mathbb{R}^d) \right\}$$

Sobolev spaces [Adams and Fournier, 2003]

A Sobolev space $H^r(\Omega)$ on a measurable set $\Omega \subset \mathbb{R}^d$ is defined by the restriction of $H^r(\mathbb{R}^d)$;

$$H^r(\Omega) := \left\{ f : \Omega \to \mathbb{R} : f = g|_\Omega,\ \exists\, g \in H^r(\mathbb{R}^d) \right\}$$

- The norm is defined by

$$\|f\|_{H^r(\Omega)} := \inf \left\{ \|g\|_{H^r(\mathbb{R}^d)} : g \in H^r(\mathbb{R}^d) \text{ s.t. } f = g|_\Omega \right\}.$$

Sobolev spaces [Adams and Fournier, 2003]

A Sobolev space $H^r(\Omega)$ on a measurable set $\Omega \subset \mathbb{R}^d$ is defined by the restriction of $H^r(\mathbb{R}^d)$;

$$H^r(\Omega) := \left\{ f : \Omega \to \mathbb{R} : f = g|_\Omega,\ \exists\, g \in H^r(\mathbb{R}^d) \right\}$$

- The norm is defined by

$$\|f\|_{H^r(\Omega)} := \inf \left\{ \|g\|_{H^r(\mathbb{R}^d)} : g \in H^r(\mathbb{R}^d) \text{ s.t. } f = g|_\Omega \right\}.$$

Sobolev spaces as RKHSs

- $H^r(\Omega)$ is norm-equivalent to RKHSs of Matérn kernels [Matèrn, 1960] and Wendland kernels [Wendland, 1995].
- Let k_r be such a kernel.

Outline

Bayesian quadrature [O'Hagan, 1991, Briol et al., 2016]

- Given $X_1, \ldots, X_n$ being fixed, weights $w_1, \ldots, w_n$ are obtained by the minimization of the worst case error:

$$e_n(P; \mathcal{H}_k) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} |P_n f - Pf|$$

Bayesian quadrature [O'Hagan, 1991, Briol et al., 2016]

- Given $X_1, \ldots, X_n$ being fixed, weights $w_1, \ldots, w_n$ are obtained by the minimization of the worst case error:

$$e_n(P; \mathcal{H}_k) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} |P_n f - Pf|$$

- Can be done by solving a linear system of size n, the resulting weights being

$$\boldsymbol{w} := (w_1, \ldots, w_n)^T = G^{-1}\boldsymbol{z} \in \mathbb{R}^n, \tag{1}$$

where

$$\begin{aligned} G &:= (k(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}, \\ \boldsymbol{z} &:= \left(\int k(X_i, x) dP(x) \right)_{i=1}^n \in \mathbb{R}^n \end{aligned}$$

Convergence rates of Bayesian quadrature

Fill distance

For a bounded set $\Omega \subset \mathbb{R}^d$ and $X^n = \{X_1, \ldots, X_n\} \subset \Omega$, the fill distance is defined by

$$h_{X^n,\Omega} := \sup_{x \in \Omega} \min_{i=1,\ldots,n} \|x - X_i\|.$$

Convergence rates of Bayesian quadrature

Fill distance

For a bounded set $\Omega \subset \mathbb{R}^d$ and $X^n = \{X_1, \ldots, X_n\} \subset \Omega$, the fill distance is defined by

$$h_{X^n,\Omega} := \sup_{x \in \Omega} \min_{i=1,\ldots,n} \|x - X_i\|.$$

- Quantifies how densely X^n covers Ω.
- e.g. $h_{X^n,\Omega} = O(n^{-1/d})$ if X^n are grid points in Ω.
- e.g. $h_{X^n,\Omega} = O_p(n^{-1/d})$ if X^n are a random sample from an appropriate proposal on Ω [Oates et al., 2016a]

Convergence rates of Bayesian quadrature

Assumptions

$\Omega \subset \mathbb{R}^d$: a bounded open set, such that an interior cone condition is satisfied, and the boundary $\partial\Omega$ is Lipschitz.

P: a probability distribution on $\mathbb{R}^d$ with a bounded density function p such that $\mathrm{supp}(P) \subset \Omega$.

k_r: a kernel whose RKHS $\mathcal{H}_{k_r}(\Omega)$ is norm-equivalent to the Sobolev space $H^r(\Omega)$ of order $r > \lfloor d/2 \rfloor$.

Convergence rates of Bayesian quadrature

Assumptions

$\Omega \subset \mathbb{R}^d$: a bounded open set, such that an interior cone condition is satisfied, and the boundary $\partial\Omega$ is Lipschitz.

P: a probability distribution on $\mathbb{R}^d$ with a bounded density function p such that $\mathrm{supp}(P) \subset \Omega$.

k_r: a kernel whose RKHS $\mathcal{H}_{k_r}(\Omega)$ is norm-equivalent to the Sobolev space $H^r(\Omega)$ of order $r > \lfloor d/2 \rfloor$.

A finite sample bound

There exist constants $C > 0$ and $h_0 > 0$ independent of X^n, such that

$$e_n(P; H^r(\Omega)) \leq C h_{X^n,\Omega}^r,$$

provided that $h_{X^n,\Omega} \leq h_0$.

Convergence rates of Bayesian quadrature

Convergence rates

- Assume $h_{X^n,\Omega} = O(n^{-\alpha})$ for some $0 < \alpha \leq 1/d$ as $n \to \infty$.
- Then we have

$$e_n(P; H^r(\Omega)) = O(n^{-\alpha r}) \quad (n \to \infty).$$

Convergence rates of Bayesian quadrature

Convergence rates

- Assume $h_{X^n,\Omega} = O(n^{-\alpha})$ for some $0 < \alpha \leq 1/d$ as $n \to \infty$.
- Then we have

$$e_n(P; H^r(\Omega)) = O(n^{-\alpha r}) \quad (n \to \infty).$$

Remarks

- If $\alpha = 1/d$ (e.g. when X^n are grid points), the rate is

$$e_n(P; H^r(\Omega)) = O(n^{-r/d}) \quad (n \to \infty).$$

- This is minimax-optimal for deterministic quadrature in $H^r(\Omega)$ [Novak, 1988].

Outline

Notation, definitions and assumptions

$\Omega \subset \mathbb{R}^d$: a (bounded) open set whose boundary is Lipschitz.

P: a probability distribution on $\mathbb{R}^d$ with $\mathrm{supp}(P) \subset \Omega$.

$e_n(P; H^r(\Omega))$: the worst case error in the Sobolev space $H^r(\Omega)$ of order $r > d/2$.

Notation, definitions and assumptions

$\Omega \subset \mathbb{R}^d$: a (bounded) open set whose boundary is Lipschitz.

P: a probability distribution on $\mathbb{R}^d$ with $\mathrm{supp}(P) \subset \Omega$.

$e_n(P; H^r(\Omega))$: the worst case error in the Sobolev space $H^r(\Omega)$ of order $r > d/2$.

$C_B^s(\Omega)$: the Banach space of functions having bounded, uniformly continuous partial derivatives up to order $s \in \mathbb{N}$.

Main result 1: the assertion

Assumptions on $\{(w_i, X_i)\}_{i=1}^n \in (\mathbb{R} \times \Omega)^n$

1. $e_n(P; H^r(\Omega)) = O(n^{-b})$ for some $b > 0$ as $n \to \infty$.
2. $\sum_{i=1}^n |w_i| = O(n^c)$ for some $c \geq 0$ as $n \to \infty$.

Main result 1: the assertion

Assumptions on $\{(w_i, X_i)\}_{i=1}^n \in (\mathbb{R} \times \Omega)^n$

1. $e_n(P; H^r(\Omega)) = O(n^{-b})$ for some $b > 0$ as $n \to \infty$.
2. $\sum_{i=1}^n |w_i| = O(n^c)$ for some $c \geq 0$ as $n \to \infty$.

Convergence rates

- For any $f \in C_B^s(\Omega) \cap H^s(\Omega)$ with $s \leq r$,

$$|P_n f - Pf| = O(n^{-bs/r + c(r-s)/r}) \quad (n \to \infty).$$

- The exponent:

$$-bs/r + c(r-s)/r.$$

Main result 1: remarks

On the assumption $\sum_{i=1}^{n} |w_i| = O(n^c)$

- The sum of (absolute) weights should not increase quickly.
- $c = 0$ holds if $\max_{i=1,\dots,n} |w_i| = O(n^{-1})$.

  (e.g. equal weight quadrature $w_1 = \cdots w_n = 1/n$, such as Quasi Monte Carlo and Kernel Herding).

Main result 1: remarks

On the assumption $\sum_{i=1}^{n} |w_i| = O(n^c)$

- The sum of (absolute) weights should not increase quickly.
- $c = 0$ holds if $\max_{i=1,\dots,n} |w_i| = O(n^{-1})$.

  (e.g. equal weight quadrature $w_1 = \cdots w_n = 1/n$, such as Quasi Monte Carlo and Kernel Herding).

On the convergence rate $O(n^{-bs/r+c(r-s)/r})$

- The best rate is achieved when $c = 0$, resulting that

$$|P_n f - Pf| = O(n^{-bs/r}) \quad (n \to \infty).$$

- Inserting $b = r/d$ (the optimal rate in $H^r(\Omega)$),

$$|P_n f - Pf| = O(n^{-s/d}) \quad (n \to \infty),$$

  which is the optimal rate in $H^s(\Omega)$ [Novak, 1988].

Main result 1: take-home messages

Adaptability to lesser smoothness

- Equal-weight quadrature rules can adaptively achieve the optimal rate in $H^s(\Omega)$ (under an additional condition $f \in C_B^s(\Omega)$).
- Do not need to know the smoothness s of f, but only its upper-bound $s \leq r$.

Main result 1: take-home messages

Adaptability to lesser smoothness

- Equal-weight quadrature rules can adaptively achieve the optimal rate in $H^s(\Omega)$ (under an additional condition $f \in C^s_B(\Omega)$).
- Do not need to know the smoothness s of f, but only its upper-bound $s \leq r$.

... but

- The result does not apply to Bayesian quadrature, for which weights are given without constraint.
- $\Rightarrow$ The main result 2.

Main result 2: a preliminary

Separation radius

For design points $X^n := \{X_1, \ldots, X_n\}$, the separation radius q_{X_n} is defined by

$$q_{X_n} := \frac{1}{2} \min_{i \neq j} \|X_i - X_j\|.$$

Main result 2: a preliminary

Separation radius

For design points $X^n := \{X_1, \ldots, X_n\}$, the separation radius q_{X_n} is defined by

$$q_{X_n} := \frac{1}{2} \min_{i \neq j} \|X_i - X_j\|.$$

- The minimum distance between distinct points.
- e.g., $q_{X^n} = \Theta(n^{-1/d})$ if X^n are grid points in Ω.
- A key quantity in stability analysis of scattered data approximation [Schaback, 1995, Wendland, 2005].

Main result 2: the assertion

Assumptions on $\{(w_i, X_i)\}_{i=1}^n \in (\mathbb{R} \times \Omega)^n$

1. $e_n(P; H^r(\Omega)) = O(n^{-b})$ for some $b > 0$ as $n \to \infty$.
2. $q_{X^n} = \Theta(n^{-a})$ for some $a > 0$ as $n \to \infty$.

Main result 2: the assertion

Assumptions on $\{(w_i, X_i)\}_{i=1}^n \in (\mathbb{R} \times \Omega)^n$

1. $e_n(P; H^r(\Omega)) = O(n^{-b})$ for some $b > 0$ as $n \to \infty$.
2. $q_{X^n} = \Theta(n^{-a})$ for some $a > 0$ as $n \to \infty$.

Convergence rates

- For any $f \in C_B^s(\Omega) \cap H^s(\Omega)$ with $s \leq r$, we have

$$|P_n f - Pf| = O(n^{-\min(b-a(r-s), as)}) \quad (n \to \infty).$$

- The best rate is obtained when $a = b/r$, resulting that

$$|P_n f - Pf| = O(n^{-bs/r}) \quad (n \to \infty).$$

Main result 2: remarks

On the assumption $q_{X^n} = \Theta(n^{-a})$

- Distinct design points should not be very close to each other.
- If $b = r/d$ (the optimal rate in $H^r(\Omega)$),

$$a = b/r = 1/d,$$

which is satisfied if, e.g., X^n are grid points.

Main result 2: remarks

On the assumption $q_{X^n} = \Theta(n^{-a})$

- Distinct design points should not be very close to each other.
- If $b = r/d$ (the optimal rate in $H^r(\Omega)$),

$$a = b/r = 1/d,$$

which is satisfied if, e.g., X^n are grid points.

On the rate $O(n^{-bs/r})$

- If $b = r/d$ (the optimal rate in $H^r(\Omega)$),

$$|P_n f - Pf| = O(n^{-s/d}) \quad (n \to \infty),$$

which is the optimal rate in $H^s(\Omega)$ [Novak, 1988].
- The same rate as in the Main result 1, but with a different assumption.

Main result 2: take-home messages

- For quadrature rules with unconstrained non-equal weights (e.g. Bayesian quadrature), distinct design points should not be very close to each other.
- The adaptation to the (unknown) lesser smoothness s also occurs (as in the Main result 1), if the above point is satisfied.

Outline

Bayesian quadrature in misspecified settings

Assumptions on design points X^n

1. $h_{X^n,\Omega} = O(n^{-\alpha})$ as $n \to \infty$ for some $0 < \alpha \leq 1/d$.
2. There exist $c_q > 0$ such that

$$h_{X^n,\Omega} \leq c_q q_{X^n}.$$

Bayesian quadrature in misspecified settings

Assumptions on design points X^n

1. $h_{X^n,\Omega} = O(n^{-\alpha})$ as $n \to \infty$ for some $0 < \alpha \leq 1/d$.
2. There exist $c_q > 0$ such that

$$h_{X^n,\Omega} \leq c_q q_{X^n}.$$

Convergence rates

- For all $f \in C_B^s(\Omega) \cap H^s(\Omega)$, we have

$$|P_n f - Pf| = O(n^{-\alpha s}) \quad (n \to \infty).$$

- The best rate is achieved when $\alpha = 1/d$,

$$|P_n f - Pf| = O(n^{-s/d}) \quad (n \to \infty).$$

Bayesian quadrature in misspecified settings

On the assumption $h_{X^n,\Omega} \leq c_q q_{X^n}$

- If this assumption is satisfied, X^n are called quasi-uniform [Wendland, 2005].
- e.g. satisfied if X^n are grid points.

Bayesian quadrature in misspecified settings

On the assumption $h_{X^n,\Omega} \leq c_q q_{X^n}$

- If this assumption is satisfied, X^n are called quasi-uniform [Wendland, 2005].
- e.g. satisfied if X^n are grid points.

On the rate $O(n^{-s/d})$

- The optimal rate for deterministic quadrature in $H^s(\Omega)$.

Outline

Conclusions

Ongoing/future work

- Other RKHSs (e.g. Gauss, tensor product Sobolev, etc.)
- Adaptive (sequential) Bayesian quadrature
- Randomized design points; preliminary results in [Kanagawa et al., 2016, Section 4]

Preprint

Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings
M. Kanagawa, B. K. Sriperumbudur and K. Fukumizu
arXiv:1709.00147 [math.NA]


Adams, R. A. and Fournier, J. J. F. (2003).
Sobolev Spaces.
Academic Press, New York, 2nd edition.


Aronszajn, N. (1950).
Theory of reproducing kernels.
Transactions of the American Mathematical Society, 68(3), pages 337–404.


Bach, F., Lacoste-Julien, S., and Obozinski, G. (2012).
On the equivalence between herding and conditional gradient algorithms.
In *Proceedings of the 29th International Conference on Machine Learning (ICML2012)*, pages 1359–1366.


Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. (2016).
Probabilistic integration: A role for statisticians in numerical analysis?
arXiv:1512.00933v4 [stat.ML].

Chen, Y., Welling, M., and Smola, A. (2010).
Supersamples from kernel-herding.
In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, pages 109–116.

Dick, J., Kuo, F. Y., and Sloan, I. H. (2013).
High dimensional numerical integration - the Quasi-Monte Carlo way.
Acta Numerica, 22(133-288).

Hickernell, F. J. (1998).
A generalized discrepancy and quadrature error bound.
Mathematics of Computation of the American Mathematical Society, 67(221):299–322.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. (2016).

Convergence guarantees for kernel-based quadrature rules in misspecified settings.
In *Advances in Neural Information Processing Systems 29*.

Matèrn, B. (1960).
Spatial variation.
Meddelanden fran Statens Skogsforskningsinstitut, 49(5).

Novak, E. (1988).
Deterministic and Stochastic Error Bounds in Numerical Analysis.
Springer-Verlag.

Oates, C. J., Cockayne, J., Briol, F.-X., and Girolami, M. (2016a).
Convergence rates for a class of estimators based on stein's identity.
arXiv:1603.03220v2 [math.ST].

Oates, C. J., Papamarkou, T., and Girolami, M. (2016b).
The controlled thermodynamic integral for Bayesian model evidence evaluation.
Journal of the American Statistical Association, 111(514):634–645.

O'Hagan, A. (1991).
Bayes–Hermite quadrature.
Journal of Statistical Planning and Inference, 29:245–260.

Schaback, R. (1995).
Error estimates and condition numbers for radial basis function interpolation.
Advances in Computational Mathematics, 3(3):251–264.

Wendland, H. (1995).
Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree.
Advances in Computational Mathematics, 4(1):389–396.

Wendland, H. (2005).
Scattered Data Approximation.
Cambridge University Press, Cambridge, UK.