Minimum Discrepancy Methods in Uncertainty Quantification

Chris. J. Oates†
Newcastle University
September 14, 2021

Contents

1 Classical Discrepancy Theory .................................................. 2

2 Numerical Cubature ................................................................... 4
2.1 Koksma–Hlawka Inequality ...................................................... 4
2.2 Cubature Error Representer .................................................... 6
2.3 Reproducing Kernel Hilbert Spaces ........................................... 7

3 Maximum Mean Discrepancy ...................................................... 10
3.1 Optimal Quantisation ............................................................ 12
3.2 Optimal Approximation ........................................................ 15

4 Stein Discrepancy .................................................................... 19
4.1 Stein Operators .................................................................... 20
4.2 Optimal Quantisation ............................................................ 22
4.3 Optimal Approximation ........................................................ 23

5 Partial Solutions to Exercises .................................................. 29
5.1 Exercise 1 ............................................................................ 29
5.2 Exercise 2 ............................................................................ 29
5.3 Exercise 3 ............................................................................ 29
5.4 Exercise 5 ............................................................................ 30

*The lectures were prepared for the École Thématique sur les Incertitudes en Calcul Scientifique (ETICS) in September 2021.
†Correspondence should be sent to chris.oates@ncl.ac.uk.
Quantisation is the act of approximating a probability distribution $P$, defined on an infinite set, with a discrete distribution $\sum_{i=1}^{n} w_i \delta(x_i)$ supported on a finite set of states $\{x_i\}_{i=1}^{n}$. Here $\delta(x_i)$ denotes a Dirac distribution centred at $x_i$. The left image can be considered to represent a probability density function (PDF) $p(x)$ for $P$, defined for $x \in [0, 1]^d$, and the dots in the right image can be considered to represent $x_i$, with the size of the dots representing $w_i$. This image is due to Gräf et al. [2012].

These lectures concern the discrete approximation of objects that are in some sense infinite-dimensional. This problem is ubiquitous to numerical computation in general. Specifically, we will consider discrete approximation of probability distributions $P$ that may be defined on an infinite set, such as $[0, 1]^d$ or $\mathbb{R}^d$. See Figure 1. The basic questions here are: (1) how many states are required to achieve a given level of approximation? (2) how can such approximations be constructed?

1 Classical Discrepancy Theory

Here we start with some motivation from a classical perspective, which considers approximation of the uniform distribution $P$ on $[0, 1]^d$ by an (un-weighted) collection of states $\{x_i\}_{i=1}^{n}$. For $x \in [0, 1]^d$, let $[0, x] = [0, x_1] \times \cdots \times [0, x_d]$.

**Definition 1** (Local discrepancy). The local discrepancy of a collection of states $x_1, \ldots, x_n \in [0, 1]^d$ at $a \in [0, 1]^d$ is

$$\Delta(a) = \Delta(a; x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[0,a]}(x_i) - \prod_{j=1}^{d} a_j.$$
Figure 2: Star discrepancy is defined as the largest difference between the volume of a hyper-rectangle \([0, \mathbf{a}]\) and the proportion of states \(x_i\) (blue dots) which are contained in the hyper-rectangle, the so-called local discrepancy at \(\mathbf{a} \in [0,1]^d\). In the left hand panel, the proportion of states in the shaded hyper-rectangle is exactly equal to the volume of the hyper-rectangle, while this is not the case in the right hand panel.

**Definition 2** (Star discrepancy). The star discrepancy of \(x_1, \ldots, x_n \in [0,1]^d\) is

\[
D^*_n = D^*_n(x_1, \ldots, x_n) = \sup_{\mathbf{a} \in [0,1]^d} |\Delta(\mathbf{a}; x_1, \ldots, x_n)|.
\]

See the illustration in Figure 2.

**Remark 1.** In dimension \(d = 1\) is is clear that a regular grid \(x_1 = 0, x_2 = 1/(n-1), \ldots, x_n = 1\) minimises \(D^*_n(x_1, \ldots, x_n)\). In this univariate setting, you may recognise that the star discrepancy from the Kolmogorov–Smirnov uniformity test.

**Remark 2.** It is simplifies discussion to anchor the hyper-rectangle at \(0\), but one can also consider an alternative to star discrepancy with hyper-rectangles \([\mathbf{a}, \mathbf{b}]\) that are un-anchored. This alternative discrepancy takes values in \([D^*_n, 2^d D^*_n]\), so in terms of fixed \(d\) asymptotics its behaviour is identical.

**Remark 3.** It may surprise you how little is known about star discrepancy. In dimensions \(d \leq 2\), it has been proven that there exist a constant \(0 < C < \infty\) such that, for any choice of \(x_1, \ldots, x_n\) and \(n \in \mathbb{N}\),

\[
C \frac{(\log n)^{d-1}}{n} \leq D^*_n(x_1, \ldots, x_n),
\]

but for dimension \(d > 2\) this bound is only conjectured.

**Remark 4.** There are numerous algorithms which aim to generate collections of states with small star discrepancy, including the Halton sequence, the Hammersley set, Sobol sequences, and various non-independent sampling methods, which are sometimes collectively referred to...
as quasi Monte Carlo (QMC) methods. For some of these methods it is known that, for an appropriate constant $0 < C < \infty$ and subsequence of values $n$ in $\mathbb{N}$,

$$D^*_n(x_1, \ldots, x_n) \leq C \frac{(\log n)^d}{n},$$

meaning that the rate of convergence in $n$ is close to the conjectured optimal rate. QMC will not be discussed further, because our aim in these lectures is to deal with more general probability distributions $P$ that arise in applications of uncertainty quantification.

**Remark 5.** A regular grid consisting of of $n = m^d$ states does not minimise star discrepancy in dimension $d > 1$. Indeed, for a regular grid (i.e. a $d$ dimensional Cartesian product of regular grids over the unit interval), one can show that

$$\frac{1}{2n^{1/d}} \leq D^*_n \leq \frac{d}{2n^{1/d}}.$$

See Leobacher and Pillichshammer [2014, Remark 2.20].

At this point it should be clear that even the simplest quantisation problems can be far from trivial. The aim of the next section is to relate the slightly abstract notion of quantisation to concrete problems of numerical integration.

**Chapter Notes** The presentation of star discrepancy followed Chapter 15 of Owen [2013], which is currently a freely available online textbook. The same reference provides an excellent introduction to QMC.

## 2 Numerical Cubature

One of the most basic operations that one could hope to perform with a probability distribution is to compute expectations of random variables; i.e. to compute integrals of the form $\int f \, dP$, or $\int f(x)p(x)\, dx$ if the probability distribution $P$ admits a probability density function (PDF) $p(x)$. In general such integrals do not possess a closed form and numerical integration (also called cubature) will be required. Quantisation is useful for cubature, since we can replace $P$ with a discrete approximation $\sum_{i=1}^n w_i \delta(x_i)$ to obtain a closed form numerical approximation $\sum_{i=1}^n w_i f(x_i)$ to the integral. Approximations of this form are sometimes called cubature rules. In this section we will see how star discrepancy can be used to analyse the accuracy of cubature rules in the case where $P$ is a uniform distribution on $[0, 1]^d$.

### 2.1 Koksma–Hlawka Inequality

Let $x_u$ denote the components of a $d$-dimensional vector $x$ that are indexed by the set $u \subseteq \{1, \ldots, d\}$. The shorthand $f(x_u, 1)$ will be used to represent $f(y)$, where the vector $y$ is defined by $y_i = x_i$ if $i \in u$ and otherwise $y_i = 1$. The mixed partial derivative of $f$ with respect to each of the co-ordinates in $u$ is denoted $\partial^{\left|u\right|} f / \partial x_u$, again for $u \subseteq \{1, \ldots, d\}$. 

4
Definition 3 (Variation). For \( f : [0, 1]^d \to \mathbb{R} \) with continuous mixed partial derivatives, we define the variation of \( f \) to be

\[
\|f\|_1 = \sum_{u \subseteq \{1, \ldots, d\}} \int_{[0,1]^{|u|}} \left| \frac{\partial^{|u|} f}{\partial x_u} (x_u, 1) \right| dx,
\]

where the sum runs over all \( 2^d \) subsets \( u \subseteq \{1, \ldots, d\} \).

The term “variation” is overloaded in the literature and we use it only informally to give a name to the norm defined in Definition 3. The concept of variation enables the accuracy of cubature rules to be analysed:

Theorem 1 (Koksma–Hlawka inequality). Let \( f : [0, 1]^d \to \mathbb{R} \) have continuous mixed partial derivatives. Then

\[
\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{[0,1]^d} f(x)dx \right| \leq \|f\|_1 D_n(x_1, \ldots, x_n). \tag{1}
\]

Remark 6. The term \( \|f\|_1 \) in (1) quantifies the complexity of the integrand, while the star discrepancy quantifies the suitability of the states \( \{x_i\}_{i=1}^n \). Thus the quality of the set \( \{x_i\}_{i=1}^n \) as a quantisation of \( P \) controls the accuracy of the cubature rule.

Remark 7. The \( f(1) \) term from \( \|f\|_1 \) can actually be removed, since the left hand side of (1) is invariant to \( f(1) \), meaning that we can apply (1) to the function \( x \mapsto f(x) - f(1) \) instead, to obtain a tighter bound. If we do that, then we obtain (up to small technicalities) the original formulation of Koksma–Hlawka.

Later we will prove Theorem 1 in full, but for now we aim to present an elementary proof of Theorem 1 in the case \( d = 1 \). For this we need the following:

Lemma 1. Let \( f : [0, 1] \to \mathbb{R} \) be continuously differentiable and let \( x_1, \ldots, x_n \in [0, 1] \). Then

\[
\frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{0}^{1} f(x)dx = \int_{0}^{1} \Delta(x) f'(x)dx
\]

where \( \Delta(x) = \Delta(x; x_1, \ldots, x_n) \) is the local discrepancy.

Proof. Note that the regularity assumption allows us to write

\[
f(x) = f(1) - \int_{x}^{1} f'(y)dy. \tag{2}
\]
Substituting (2) into the expression for the cubature error, we obtain
\[
\frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{0}^{1} f(x) \, dx = \int_{0}^{1} \int_{x_i}^{1} f'(y) \, dy \, dx - \frac{1}{n} \sum_{i=1}^{n} \int_{x_i}^{1} f'(y) \, dy
\]
\[
= \int_{0}^{1} \int_{0}^{y} f'(y) \, dy \, dx - \frac{1}{n} \sum_{i=1}^{n} \int_{x_i}^{1} f'(y) \, dy
\]
\[
= \int_{0}^{1} f'(y) \left[ y - \frac{1}{n} \sum_{i=1}^{n} 1_{(x_i, 1]}(y) \right] \, dy = -\Delta(y)
\]
as required.

\[\square\]

**Proof of Theorem 1 (d = 1).** From Lemma 1 we have
\[
\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{0}^{1} f(x) \, dx \right| \leq \int_{0}^{1} |\Delta(x)f'(x)| \, dx \leq \sup_{x \in [0,1]} |\Delta(x)| \int_{0}^{1} |f'(x)| \, dx
\]
\[= D_n^* \|f\|_1 - |f(1)|\]
as required.

\[\square\]

**Chapter Notes**  See Remark 2.19 in [Dick and Pillichshammer, 2010] for a detailed discussion of how Theorem 1 relates to the original formulation of Koksma and Hlawka. Lemma 1 and the proof of Theorem 1 for \(d = 1\) can be found in Section 2.2 in [Dick and Pillichshammer, 2010]; see also Theorem 15.1 in [Owen, 2013]. The one-dimensional version of the Koksma–Hlawka inequality is sometimes called Koksma’s inequality or Zaremba’s identity [Dick and Pillichshammer, 2010, p18].

### 2.2 Cubature Error Representer

In this section and the next, we will introduce the mathematical tools that are needed to prove Theorem 1 in full. These tools will also be useful later, when we consider practical algorithms for quantisation of general probability distributions \(P\). The aim is to generalise the concept of variation in Definition 3 to allow for functions \(f\) of different regularity to be integrated.

The basic idea is as follows: we consider the set \(S(k)\) of all functions of the form \(f(x) = \sum_{i=1}^{m} b_i k(x, y_i)\), where \(k\) is to be specified, the \(y_i\) are fixed states, and \(n \in \mathbb{N}\). The function \(k\) determines the regularity of the elements in \(S(k)\); for example, if \(k(x, y) = \exp(-\|x - y\|)\) then the elements of \(S(k)\) are continuous but not differentiable, while if \(k(x, y) = \exp(-\|x - y\|^2)\) then the elements of \(S(k)\) are infinitely differentiable. See Figure 3. Since we aim to perform mathematical analysis, we will want to endow the set \(S(k)\) with mathematical structure that we can exploit. It is clearly a vector space (over the reals) of functions when
(pointwise) addition and scalar multiplication are defined. In addition to that, we will want to make use of an inner product

\[ \langle f, g \rangle_{S(k)} = \sum_{i=1}^{m} \sum_{j=1}^{n} b_ic_jk(y_i, z_j), \quad f(x) = \sum_{i=1}^{m} b_ik(x, y_i), \quad g(x) = \sum_{j=1}^{n} c_jk(x, z_j), \]

for which we must require that \( k \) is symmetric (i.e. \( \langle f, g \rangle_{S(k)} = \langle g, f \rangle_{S(k)} \)) and positive definite (i.e. \( \langle f, f \rangle_{S(k)} > 0 \) for all \( f \neq 0 \)). This inner product is useful because it satisfies a reproducing property, meaning that

\[ \langle f, k(\cdot, x) \rangle_{S(k)} = f(x), \]

and suggesting the formal manipulation

\[
\frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{[0,1]^d} f(x) \, dx = \frac{1}{n} \sum_{i=1}^{n} \langle f, k(\cdot, x_i) \rangle_{S(k)} - \int_{[0,1]^d} \langle f, k(\cdot, y) \rangle_{S(k)} \, dy
\]

\[
\overset{?}{=} \left\langle \left[ f, \frac{1}{n} \sum_{i=1}^{n} k(\cdot, x_i) - \int_{[0,1]^d} k(\cdot, y) \, dy \right] \right\rangle_{S(k)},
\]

where \( e(\cdot) \) is referred to as the representer of the cubature error. The representer \( e(\cdot) \) would completely characterise the error of the (un-weighted) cubature rule based on the states \( \{x_i\}_{i=1}^{n} \). For example, if \( e(x) = 0 \) for all \( x \) then the cubature rule would be exact for all integrands \( f \in S(k) \).

Unfortunately the inner product space \( S(k) \) is not complete, in the sense that limits of functions of the form \( f(x) = \sum_{i=1}^{m} b_ik(x, y_i) \) need not be elements of the set \( S(k) \). In particular, the integral \( \int_{[0,1]^d} k(\cdot, y) \, dy \) is not an element of \( S(k) \), meaning that the formal manipulation above is not well-defined. For this technical reason we work with a larger set \( H(k) \), called the completion of \( S(k) \), whose definition we present next.

### 2.3 Reproducing Kernel Hilbert Spaces

The main mathematical tool that we will exploit is that of a reproducing kernel:

**Definition 4** (Reproducing kernel Hilbert space). Let \( \mathcal{X} \) be a set and consider a symmetric and positive definite function \( k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \). Then a reproducing kernel Hilbert space (RKHS) with reproducing kernel (or simply kernel) \( k \) is an inner product space \( H(k) \) of functions \( f: \mathcal{X} \rightarrow \mathbb{R} \), such that

1. \( k(\cdot, x) \in H(k) \) for all \( x \in \mathcal{X} \)
2. \( \langle f, k(\cdot, x) \rangle_{H(k)} = f(x) \) for all \( x \in \mathcal{X} \) and all \( f \in H(k) \).
Figure 3: Reproducing kernel Hilbert spaces (RKHSs): The left panel represents elements from a reproducing kernel Hilbert space (RKHS) whose kernel is non-differentiable, while the right panel corresponds to an infinitely differentiable kernel.

**Remark 8.** Given a symmetric positive definite function $k$, it can be shown that there exists a unique RKHS $\mathcal{H}(k)$. Conversely, each RKHS admits a unique reproducing kernel, and that kernel is symmetric and positive definite.

In general it is difficult to characterise the inner product induced by a reproducing kernel, and hence the elements of the RKHS. However, there are a number of important cases where this can be carried out:

**Example 1.** The linear span of a finite collection of functions $e_1(x), \ldots, e_p(x)$ can be endowed with the structure of an RKHS with reproducing kernel $k(x, y) = \sum_{i=1}^{p} e_i(x) e_i(y)$. The induced inner product is $\langle f, g \rangle_{\mathcal{H}(k)} = \sum_{i=1}^{n} b_i c_i$, where $f(x) = \sum_{i=1}^{n} b_i e_i(x)$ and $g(x) = \sum_{i=1}^{n} c_i e_i(x)$.

**Example 2.** The kernel

$$k(x, y) = \prod_{i=1}^{d} (1 + \min(1 - x_i, 1 - y_i)), \quad x, y \in [0, 1]^d$$

reproduces a Hilbert space with inner product

$$\langle f, g \rangle_{\mathcal{H}(k)} = \sum_{u \subseteq \{1, \ldots, d\}} \int_{[0, 1]^s} \frac{\partial^{\mid u \mid} f}{\partial x_u}(x_u, 1) \frac{\partial^{\mid u \mid} g}{\partial x_u}(x_u, 1) dx_u.$$

**Standing Assumption 1.** For all reproducing kernels $k$ considered in the sequel, we assume that $(x, y) \mapsto k(x, y)$ is a continuous function.

**Remark 9.** Identical manipulation to that presented in Section 2.2 shows that the (Riesz) representer of the cubature error is

$$e(x) = e(x; x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} k(x, x_i) - \int_{[0, 1]^d} k(x, y) dy.$$
Note that the integral of the kernel is well-defined from Standing Assumption 1. The interchange of integral and inner product in Section 2.2 requires justification; this will be provided later in Lemma 2.

Armed with reproducing kernels and the cubature error representer, we can now prove Theorem 1 in full. In what follows we let $x_{i,j}$ denote the $j$th coordinate of the vector $x_i$ and we let $x_{i,u}$ denote the components of the vector $x_i$ that are indexed by the set $u \subseteq \{1, \ldots, d\}$.

**Proof of Theorem 1.** Consider the kernel $k$ in (3):

$$k(x, y) = \prod_{i=1}^{d} (1 + \min(1 - x_i, 1 - y_i))$$

and compute the cubature error representer

$$e(x) = \frac{1}{n} \sum_{i=1}^{n} k(x, x_i) - \int_{[0,1]^d} k(x, y) dy$$

$$= \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} (1 + \min(1 - x_i, 1 - x_{i,j})) - \prod_{i=1}^{d} \frac{3 - x_i^2}{2}.$$ 

Then, for $u \subseteq \{1, \ldots, d\}$ and $(x_u, 1) \notin \{x_1, \ldots, x_n\}$,

$$\frac{\partial |u|e}{\partial x_u}(x_u, 1) = (-1)^{|u|} \left( \frac{1}{n} \sum_{i=1}^{n} 1_{[0,1]^d}(x_{i,u}) (x_{i,u}) - \prod_{i \in u} x_i \right).$$

The assumed regularity of $f$ ensures that $f \in \mathcal{H}(k)$. Plugging (5) into (4), we obtain

$$\frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{[0,1]^d} f(x) dx = \langle f, e \rangle_{\mathcal{H}(k)}$$

$$= \sum_{u \subseteq \{1, \ldots, d\}} (-1)^{|u|} \int_{[0,1]^d} \frac{\partial |u|f}{\partial x_u}(x_u, 1) \Delta(x_u, 1) dx_u.$$ 

Finally, taking absolute values and using the bound $|\Delta(x_u, 1)| \leq \sup_{x \in [0,1]^d} |\Delta(x)| = D_n^*$, we arrive at [1].

**Chapter Notes** There are several excellent introductions to the theory of reproducing kernels, including Wendland [2004], Berlinet and Thomas-Agnan [2011]. The presentation of the Koksma–Hlawka inequality from a reproducing kernel perspective follows Section 2.4 in Dick and Pillichshammer [2010]. A proof of the Koksma–Hlawka inequality, which does not require reproducing kernels, can be found as Theorem 5.5 in Chapter 2 of Kuipers and Niederreiter [1974]. Example 2 can be found in Section 2.4 of Dick and Pillichshammer [2010].

9
Figure 4: Kernel mean embedding: Two probability distributions $P$ and $Q$ are mapped to their respective elements $\mu_P$ and $\mu_Q$ in the RKHS $\mathcal{H}(k)$. The distance (in $\mathcal{H}(k)$) between these kernel mean embeddings $\mu_P$ and $\mu_Q$ is called the maximum mean discrepancy (MMD) between $P$ and $Q$.

3 Maximum Mean Discrepancy

Now we move beyond the uniform distribution $P = \mathcal{U}([0, 1]^d)$ and consider general probability distributions $P$ on general domains $\mathcal{X}$. The aim is to present a modern treatment of discrepancy and cubature error, that generalises the previous results.

**Definition 5** (Kernel mean embedding). For a kernel $k$ and a probability distribution $P$, we call $\mu_P = \int k(\cdot, x) dP(x)$ the kernel mean embedding of $P$ in $\mathcal{H}(k)$, whenever it is well-defined (see Lemma 2).

**Lemma 2.** If $\int \sqrt{k(x, x)} dP(x) < \infty$ then $\mu_P(x) \in \mathcal{H}(k)$ and $\int f(x) dP(x) = \langle f, \mu_P \rangle_{\mathcal{H}(k)}$.

**Proof.** Consider the linear operator $Lf = \int f(x) dP(x)$. Then

$$|Lf| = \left| \int f(x) dP(x) \right| \leq \int |f(x)| dP(x) \quad (6)$$

$$= \int |\langle f, k(\cdot, x) \rangle_{\mathcal{H}(k)}| dP(x) \quad (7)$$

$$\leq \int \|f\|_{\mathcal{H}(k)} \|k(\cdot, x)\|_{\mathcal{H}(k)} dP(x) \quad (8)$$

$$= \int \sqrt{k(x, x)} dP(x) \|f\|_{\mathcal{H}(k)} \quad (9)$$

1These lecture notes deliberately avoid discussion of measure theory, but to be fully rigorous we restrict attention to Borel measures $P$ defined on a topological space $\mathcal{X}$. 

10
where (6) is Jensen’s inequality, (7) is the reproducing property, and (8) is Cauchy–Schwarz. This shows that $L$ is a \textit{bounded linear operator} from $\mathcal{H}(k)$ to $\mathbb{R}$. Thus, from the Riesz representation theorem, there exists $h \in \mathcal{H}(k)$ such that $L f = \langle f, h \rangle_{\mathcal{H}(k)}$. Taking $f(x) = k(y, x)$ and using the reproducing property leads to $\int k(y, x)dP(x) = Lf = \langle f, h \rangle_{\mathcal{H}(k)} = h(y)$, so that $h = \int k(\cdot, x)dP(x)$, and so $\mu_p = h \in \mathcal{H}(k)$ with $Lf = \langle f, \mu_P \rangle_{\mathcal{H}(k)}$, as was claimed. 

\textbf{Standing Assumption 2.} For all reproducing kernels $k$ and probability distributions $P$ considered in the sequel, we assume that
\[
\int \sqrt{k(x, x)}dP(x) < \infty.
\]

In this more general setting the Riesz representer of the cubature error is the difference $e = \mu_{Q_n} - \mu_P$ of two kernel mean embeddings, where $Q_n = \sum_{i=1}^n w_i \delta(x_i)$ is the discrete distribution on which the cubature rule is based, i.e.
\[
\sum_{i=1}^n w_i f(x_i) - \int f dP = \langle f, \mu_{Q_n} - \mu_P \rangle_{\mathcal{H}(k)}, \quad \mu_{Q_n} = \sum_{i=1}^n w_i k(\cdot, x_i), \quad \mu_P = \int k(\cdot, x)dP(x).
\]

There are several different ways to systematically assess the performance of a cubature rule, but here we focus on a \textit{worst case} assessment:

\textbf{Definition 6 (Maximum mean discrepancy).} The \textit{maximum mean discrepancy} (MMD) between two distributions $P$ and $Q$ is
\[
D_k(P, Q) = \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \left| \int f dP - \int f dQ \right|,
\]
also called the \textit{worst case cubature error} in the unit ball of $\mathcal{H}(k)$.

A similar argument to Remark 9 shows that:

\textbf{Lemma 3.} $D_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}(k)}$.

\textit{Proof.} Since $e = \mu_P - \mu_Q$ is the Riesz representer of the integral approximation error, we may apply Cauchy–Schwarz to obtain
\[
\left| \int f dP - \int f dQ \right| = |\langle f, e \rangle_{\mathcal{H}(k)}| \leq \|f\|_{\mathcal{H}(k)} \|e\|_{\mathcal{H}(k)},
\]
which shows that
\[
0 \leq D_k(P, Q) = \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \left| \int f dP - \int f dQ \right| \leq \|e\|_{\mathcal{H}(k)}.
\]

If $\|e\|_{\mathcal{H}(k)} = 0$ then the bound is necessarily an equality. If not, consider $f = e/\|e\|_{\mathcal{H}(k)}$, which satisfies $\|f\|_{\mathcal{H}(k)} \leq 1$ and $\langle f, e \rangle_{\mathcal{H}(k)} = \|e\|_{\mathcal{H}(k)}$, showing that the bound is in fact an equality. 

\[11\]
This result is summarised visually in Figure 4.

If \( D_k(P, Q_n) = 0 \), the cubature rule based on \( Q_n \) will be exact for all integrands \( f \in \mathcal{H}(k) \). Does this mean that \( Q_n \) and \( P \) are identical?

**Definition 7** (Characteristic kernel). A kernel \( k \) is said to be characteristic if \( D_k(P, Q) = 0 \) implies \( P = Q \).

**Example 3** (Polynomial kernel is not characteristic). From Example 1, the kernel \( k(x, y) = \sum_{i=1}^{p} x^i y^i \) reproduces an RKHS whose elements are the polynomials of degree at most \( p \) on the domain \( \mathcal{X} = \mathbb{R} \). Thus \( D_k(P, Q) = 0 \) if and only if the moments \( \int x^i dP(x) \) and \( \int x^i dQ(x) \) are identical for \( i = 1, \ldots, p \). In particular, \( k \) is not a characteristic kernel.

**Example 4.** The Gaussian kernel \( k(x, y) = \exp(-\|x - y\|^2) \) is a characteristic kernel on \( \mathcal{X} = \mathbb{R}^d \).

The characteristic property is desirable but, on its own, it does not provide strong justification for using \( D_k(P, Q) \) to measure the discrepancy between \( P \) and \( Q \). For this reason we now introduce a stronger property, called weak convergence control. Let \( Q_n \Rightarrow P \) denote that the sequence \( (Q_n)_{n=1}^{\infty} \) converges weakly (or in distribution) to \( P \) (i.e. \( \int f dQ_n \to \int f dP \) for all functions \( f \) which are continuous and bounded).

**Definition 8** (Weak convergence control). A kernel \( k \) is said to have weak convergence control if \( D_k(P, Q_n) \to 0 \) implies that \( Q_n \Rightarrow P \).

**Remark 10.** Perhaps surprisingly, for a compact Hausdorff space \( \mathcal{X} \), a bounded\(^2\) characteristic kernel \( k \) is guaranteed to have weak convergence control. This equivalence no longer holds when the domain \( \mathcal{X} \) is non-compact, and a bounded and characteristic kernel can fail to have weak convergence control; see [Simon-Gabriel et al. 2020]. Clearly a kernel that is not characteristic fails to have weak convergence control.

Convergence control justifies attempting to minimise MMD for the purposes of quantisation and more general approximation, as we will attempt in the sequel.

**Example 5.** The Gaussian kernel \( k(x, y) = \exp(-\|x - y\|^2) \) controls weak convergence of probability distributions on \( \mathcal{X} = [0, 1]^d \). It can also be shown that the Gaussian kernel controls weak convergence on \( \mathcal{X} = \mathbb{R}^d \); this can be deduced from e.g. Theorem 7 of [Simon-Gabriel et al. 2020] and the general results in [Sriperumbudur et al. 2011].

### 3.1 Optimal Quantisation

The goal of quantisation is to find \( Q \) of the form \( Q_n = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i) \) such that \( Q_n \approx P \) in some sense, and in these lectures that sense will be MMD. To start to move toward practical
Figure 5: Monte Carlo: Independent samples (blue circles) from a “horseshoe” distribution $P$ (grey contours).

Now, considering for example the term $\langle \mu_P, \mu_Q \rangle_{\mathcal{H}(k)}$, we have

$$\langle \mu_P, \mu_Q \rangle_{\mathcal{H}(k)} = \left\langle \int k(\cdot, x) dP(x), \int k(\cdot, y) dQ(y) \right\rangle_{\mathcal{H}(k)}$$

$$= \int \int k(x, y) dP(x) dQ(y).$$

Here we have used the reproducing property, as well as using Lemma 2 to justify the exchanges of integral and inner product. Proceeding similarly with all three terms results in the expression

$$D_k(P, Q)^2 = \int \int k(x, y) dP(x) dP(y) - 2 \int \int k(x, y) dP(x) dQ(y) + \int \int k(x, y) dQ(x) dQ(y).$$

As a simple baseline method for quantisation we consider Monte Carlo (Figure 5):

**Proposition 1** (MMD of Monte Carlo). Let $x_1, \ldots, x_n \sim P$ be independent. Assume that $C := \int k(x, x) dP(x) < \infty$. Then

$$E \left[ D_k(P_n)^2 \right] \leq \frac{C}{n}.$$
Proof. From the above discussion, with \( Q = Q_n = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i) \) we obtain that

\[
D_k(P, Q_n)^2 = \iint k(x, y) dP(x) dP(y) - \frac{2}{n} \sum_{i=1}^{n} \int k(x, x_i) dP(x) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j)
\]

Taking expectations of both sides gives

\[
\mathbb{E} \left[ D_k(P, Q_n)^2 \right] = \mathbb{E} \left[ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) \right] - \iint k(x, y) dP(x) dP(y)
\]

\[
= \mathbb{E} \left[ \frac{1}{n^2} \sum_{i=1}^{n} k(x_i, x_i) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} k(x_i, x_j) \right] - \iint k(x, y) dP(x) dP(y)
\]

\[
= \mathbb{E} \left[ \frac{1}{n^2} \sum_{i=1}^{n} k(x_i, x_i) \right] - \frac{1}{n} \iint k(x, y) dP(x) dP(y)
\]

\[
\leq \frac{1}{n} \int k(x, x) dP(x)
\]

since \( x_i \sim P \) are independent.

Thus Monte Carlo sampling provides a consistent but potentially far from optimal quantisation of \( P \). Note that the convergence rate in Proposition 1 does not depend on the kernel \( k \), which highlights the inefficiency of the Monte Carlo method in this context (e.g. compare against the later Theorem 2). The goal of optimal quantisation is to quantise \( P \) using as few states \( x_i \) as possible (for a given approximation quality). A conceptually simple approach to optimal quantisation is illustrated in Figure 6, and you are encouraged to try this out:

**Exercise 1** (Optimal quantisation with MMD). Consider \( P = \mathcal{N}(0, I) \) and \( k(x, y) = \exp(-\|x - y\|^2/\sigma^2) \) on \( \mathcal{X} = \mathbb{R}^d \). (You may wish to focus on \( d = 1 \) or \( d = 2 \).)

(a) Calculate (analytically) the kernel mean embedding \( \mu_P(x) = \int k(x, y) dP(y) \).

(b) For a fixed value of \( n \) (e.g. \( n = 10 \)) and a fixed value of \( \sigma \) (e.g. \( \sigma = 1 \)), try to numerically optimise the locations of the states \( x_1, \ldots, x_n \) in order to minimise \( D_k(P, \frac{1}{n} \sum_{i=1}^{n} \delta(x_i)) \).

(c) What effect does varying the bandwidth parameter \( \sigma \) have on the approximations that are produced?

After performing Exercise 1 it should be clear that (1) MMD provides a coherent framework for optimal quantisation, but (2) direct/naive numerical optimisation of MMD may be impractical, or at least not straightforward. A neat solution is proposed next.
Figure 6: Optimal (un-weighted) quantisation: Sequential (greedy) minimisation of MMD to select $n = 8$ states $\{x_i\}_{i=1}^n$ in an approximation $Q_n$ to $P$. The numbers indicate the order in which the states $x_i$ were selected.

3.2 Optimal Approximation

Motivated by the difficulty of multivariate optimisation in $\mathcal{X} \times \cdots \times \mathcal{X}$ in Exercise 1, in this section we return to the case of independently sampled $x_i \sim P$ but now we allow for weighted point sets; i.e. approximations of the form

$$Q_n = \sum_{i=1}^n w_i \delta(x_i)$$

for some weights $w_1, \ldots, w_n \in \mathbb{R}$.

**Lemma 4.** Let $x_1, \ldots, x_n \in \mathcal{X}$ be distinct. The optimal weights

$$\arg \min_{w \in \mathbb{R}^n} D_k(P, \sum_{i=1}^n w_i \delta(x_i))$$

are the solution of the linear system

$$Kw = z$$

(9)

where $K_{ij} = k(x_i, x_j)$ and $z_i = \mu_P(x_i)$.

**Proof.** The MMD between $P$ and $Q_n = \sum_{i=1}^n w_i \delta(x_i)$ can be expressed as

$$D_k(P, Q_n)^2 = \int \int k(x, y) dP(x) dP(y) - 2 \int \int k(x, y) dP(x) dQ(y) + \int \int k(x, y) dQ(x) dQ(y)$$

$$= C - 2z^T w + w^T Kw$$

where $C = \int \int k(x, y) dP(x) dP(y)$ is independent of $w$. This is a non-degenerate quadratic form in $w$ (since $K$ is a positive definite matrix), from which the result is easily verified (e.g. differentiate w.r.t. $w$ and solve for the unique critical point). \qed
Figure 7: *Optimally weighted Monte Carlo samples:* The weights $w_1, \ldots, w_n$ are obtained by minimising MMD in the manner of Lemma 4. Blue indicates states $x_i$ with positive weights $w_i > 0$, while red indicates negative weights $w_i < 0$. The size of the circles is proportional to $|w_i|$.

See the illustration in Figure 7.

**Remark 11.** The linear system in (9), defining the optimal weights, can be numerically ill-conditioned (i.e. close to singular) when $n$ is large or when two of the $x_i$ are close together (where the terms “large” and “close” depend on the specific kernel $k$ that is used). (Of course, if $x_i$ and $x_j$ are identical we may assign $w_i = 0$ without loss of generality.) Although we will not discuss this point further, there are several techniques for numerical regularisation could be used.

**Remark 12.** In general the optimal weights can be negative and will not sum to 1. Both constraints can be enforced if required and, essentially, the results that we discuss continue to hold. See e.g. Section 2.3 of [Karvonen et al. 2018].

Our aim in the remainder of this section is to show that the optimal weights in Lemma 4 enable faster rates of convergence than the Monte Carlo rate in Proposition 1. To make this concrete we consider a particularly natural class of RKHS, introduced next. The *multi-index* notation

$$
\partial^\alpha f : \mathbf{x} \mapsto \frac{\partial^\alpha f(\mathbf{x})}{\partial \mathbf{x}^\alpha} = \frac{\partial^{||\alpha||_1}}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} f(\mathbf{x}), \quad \alpha \in \mathbb{N}_0^d
$$

will be used, and we let $|\alpha| = \alpha_1 + \cdots + \alpha_d$.

**Definition 9.** For $s > d/2$ and (sufficiently regular) $\mathcal{X} \subset \mathbb{R}^d$, the (order $s$) Sobolev space $H^s(\mathcal{X})$ is defined to be the set of functions $f : \mathcal{X} \to \mathbb{R}$ whose mixed partial derivatives $\partial^\alpha f$, $|\alpha| \leq s$, exist in $L^2(\mathcal{X})$. This becomes a Hilbert space with inner product

$$
\langle f, g \rangle_{H^s(\mathcal{X})} = \sum_{|\alpha| \leq s} \int \frac{\partial^\alpha f(\mathbf{x})}{\partial \mathbf{x}^\alpha} \frac{\partial^\alpha g(\mathbf{x})}{\partial \mathbf{x}^\alpha} d\mathbf{x}.
$$
A kernel \( k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) is called a Sobolev kernel if there exists \( 0 < c_1 < c_2 < \infty \) such that, for all \( f \in \mathcal{H}(k) \), we have \( c_1\|f\|_{H^s(\mathcal{X})} \leq \|f\|_{\mathcal{H}(k)} \leq c_2\|f\|_{H^s(\mathcal{X})} \).

**Example 6.** Let \( z^n \) denote \( \max(0, z)^n \) in shorthand. Then examples of Sobolev kernels on (sufficiently regular) \( \mathcal{X} \subseteq \mathbb{R}^d \) include the following, due to [Wendland 1998]:

| \( k(x, y) \) | (\( r = |x - y|, \ x, y \in \mathbb{R}^d \)) | order |
|----------------|-------------------------------------------------|------|
| \( (1 - r)_+ \) | \( s = 1 \) |
| \( (1 - r)^{3r+1}_+ \) | \( s = 2 \) |
| \( (1 - r)^{8r^2+5r+1}_+ \) | \( s = 3 \) |

These kernels are convenient for numerical reasons, due to their compact support, which renders \( K \) a sparse matrix.

The above inner product should be contrasted with that in Example 2. Here we are considering mixed partial derivatives of order at most \( s \), where each coordinate can in principle be differentiated more than once, whereas in Example 2 we consider mixed partial derivatives of up to order \( d \), provided that each coordinate of \( x \) is differentiated at most once.

Given the slow convergence of MMD with uniformly-weighted random samples in Proposition 1, the following is possibly very surprising:

**Theorem 2.** Let \( x_1, \ldots, x_n \sim P \) be independent and let \( w = w(x_1, \ldots, x_n) \) denote optimal weights in the sense of Lemma 4. Let \( k(x, y) \) be an (order \( s \)) Sobolev kernel. Then, under regularity conditions on the domain \( \mathcal{X} \), which is of dimension \( d \), and the distribution \( P \), there exists a constant \( 0 < C < \infty \) such that

\[
\mathbb{E} \left[ D_k \left( P, \sum_{i=1}^{n} w_i \delta(x_i) \right) \right] \leq \left( \frac{C \log(n)}{n} \right)^{s/d}.
\]

**Proof.** The following is a sketch: For simplicity assume that \( \mathcal{H}(k) \) and \( H^s(\mathcal{X}) \) have an identical inner product. Under regularity conditions on the domain \( \mathcal{X} \), which amounts to \( \mathcal{X} \) being bounded and satisfying an interior cone condition, one can obtain a sampling inequality of the form

\[
\|f - f_n\|_\infty \leq Ch_n^{s/d}\|f\|_{H^s(\mathcal{X})}
\]

where \( f_n(x) = \sum_{i=1}^{n} c_i k(x, x_i) \), \( c = K^{-1}f \), is an interpolant of the function \( f \) at the locations \( x_1, \ldots, x_n \), and \( h_n \) is the fill distance

\[
h_n = \sup_{x \in \mathcal{X}} \min_{i=1, \ldots, n} \|x - x_i\|,
\]

see e.g. Theorem 11.13 of [Wendland 2004]. Then observe that \( \int f_n dP = \sum_{i=1}^{n} c_i z_i = f^T K^{-1} z = f^T w = \sum_{i=1}^{n} w_i f(x_i) \). Thus \( \left| \sum_{i=1}^{n} w_i f(x_i) - \int f dP \right| = \left| \int f_n dP - \int f dP \right| \leq \|f_n - f\|_\infty \leq Ch_n^{s/d}\|f\|_{H^s(\mathcal{X})} \). From the definition of MMD it follows that \( D_k(P, \sum_{i=1}^{n} w_i \delta(x_i)) \leq Ch_n^{s/d} \). Under regularity conditions on \( P \), which amounts to \( P \) admitting a PDF bounded away from 0 and \( \infty \) on \( \mathcal{X} \), one can show that \( \mathbb{E}[h_n^{s/d}] \) decreases at the advertised \( O(((\log n)/n)^{s/d}) \) rate; see [Reznikov and Saff 2016].
Thus for $s = d/2$ we recover the same rate as Proposition 1 for un-weighted Monte Carlo (up to log factors), while for $s > d/2$ we obtain faster convergence in MMD.

**Remark 13.** Stronger concentration inequalities than Theorem 2 can also be established; see e.g. [Ehler et al. 2019].

So surely it is a good idea to employ optimal weights? Not necessarily - the computational cost is $\mathcal{O}(n^3)$ in general and numerical ill-conditioning requires careful treatment. In QMC an important goal is to find a deterministic sequence of sets of (un-weighted) states whose computation is $\mathcal{O}(n)$, such that MMD is asymptotically minimised. Thus, at least for the simple forms of $P$ for which a QMC method has been discovered, the QMC approach would usually be preferred. The RKHS/MMD perspective on QMC was popularised by [Hickernell 1998].

**Exercise 2** (Rates of convergence and MMD). Consider the uniform distribution $P = \mathcal{U}([0, 1])$ together with the Sobolev kernels $k$ of orders $s \in \{1, 2, 3\}$ in Example 6.

(a) Calculate the kernel mean embeddings $\mu_P(x) = \int_0^1 k(x, y) dy$.

(b) Generate and store a sequence $(x_i)_{i=1}^{100}$ of independent samples from $P$.

(c) For each $n = 1, \ldots, 100$ and each $s = 1, 2, 3$, calculate and plot the values of

$$D_k \left( P, \frac{1}{n} \sum_{i=1}^n \delta(x_i) \right) \quad \text{and} \quad D_k \left( P, \sum_{i=1}^n w_i^{(n)} \delta(x_i) \right)$$

where $w_{(m)} = (w_1^{(m)}, \ldots, w_n^{(m)})$ are the optimal weights obtained by solving the $m$-dimensional linear system in Lemma 4 based on the states $\{x_i\}_{i=1}^m$.

(d) What rates of convergence would you expect to observe for these quantities as $n \to \infty$, and do your experiments agree with these rates of convergence?

**Chapter Notes** The presentation of Lemma 2 follows Lemma 3.1 of [Muandet et al. 2016]. Lemma 2 presented an elementary argument for why kernel mean embeddings are well-defined, but a more general framework is the Bochner integral. Bochner’s criterion for integrability states that a Bochner-measurable function $F : [0, 1]^d \to \mathcal{H}(k)$ is Bochner integrable if and only if $\int \|F(x)\|_{\mathcal{H}(k)} dP(x) < \infty$. Here we take $F(x) = k(\cdot, x)$, noting that $\|k(\cdot, x)\|_{\mathcal{H}(k)} = \sqrt{k(x, x)}$, which recovers the condition in Lemma 2. There is an elegant criterion to determine when a translation-invariant kernel $k$ (i.e. $k(x, y) = \phi(x - y)$ for some function $\phi$) is characteristic; see Section 2.1 of [Muandet et al. 2016]. Stronger concentration inequalities than Proposition 1 for Monte Carlo MMD have been established; see Section 3.3 of [Muandet et al. 2016]. The case of Theorem 2 where $\mathcal{X}$ is a smooth, connected, closed Riemannian manifold of dimension $d$ is presented in [Ehler et al. 2019]. (This requires some generalisation of the definition of a Sobolev kernel.) The fastest known rates for explicit
constructions of weighted approximations in the case of Exercise 1 are (at the time of writing) due to Karvonen et al. [2021]. Greedy optimisation can provide a practical solution to optimal quantisation problems like Exercise 1 [Pronzato and Zhigljavsky, 2020, Teymur et al., 2021], as can the sophisticated numerical methods for high-dimensional optimisation used to produce Figure 1 [Gräf et al., 2012]. Several tricks are available to reduce the cost of solving the linear system in (9); see e.g. Karvonen and Särkkä [2018].

4 Stein Discrepancy

In this final section we aim to perform optimal quantisation of a distribution \( P \) that admits a PDF \( p(x) \) on \( x \in \mathbb{R}^d \), such that

\[
p(x) = \frac{\tilde{p}(x)}{Z},
\]

where \( \tilde{p} \) can be exactly evaluated but \( Z \), and hence \( p(x) \), cannot easily be evaluated or even approximated. This setting is typical in applications of Bayesian inference, where we have

\[
p(x) = \frac{\pi(x) \mathcal{L}(x)}{Z}
\]

where \( \pi(x) \) is a prior PDF, \( \mathcal{L}(x) \) is a likelihood, and the implicitly defined normalisation constant \( Z \) is the marginal likelihood. The integral

\[
Z = \int \pi(x) \mathcal{L}(x) dx
\]

is often extremely challenging to evaluate due to localised regions in which \( \mathcal{L} \) takes very large values. Several methods have been developed in the statistics, applied probability, physics and machine learning literatures to approximate distributions \( P \) with these characteristics, including Markov chain Monte Carlo (MCMC), sequential Monte Carlo (SMC), and variational inference. These techniques do not typically attempt optimal quantisation, since even the basic quantisation task can be difficult.

The aim of this section is to discuss whether the techniques described in Section 3 can be applied in this more challenging context. The apparent difficulty is that we cannot compute integrals with respect to \( P \), such as \( \int k(\cdot, x) dP(x) \), which are required for computation of MMD. A hint at a possible solution is provided by the following result:

**Lemma 5.** Suppose \( k_P : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) is a symmetric positive definite kernel with \( \int k_P(\cdot, x) dP(x) = 0 \) for all \( x \in \mathcal{X} \). Then

\[
D_{k_P}(Q) = D_{k_P}(P, Q) = \sup_{\|f\|_{\mathcal{H}(k_P)} \leq 1} \left| \int f dQ \right|.
\]

**Proof.** For all \( f \in \mathcal{H}(k_P) \) it holds that \( \int f dP = 0 \), whence the result. Indeed, from the reproducing property, and using Lemma 2 with Standing Assumption 2 to interchange integral with inner product, \( \int f dP = \int \langle f, k(\cdot, x) \rangle_{\mathcal{H}(k_P)} dP(x) = \langle f, \int k(\cdot, x) dP(x) \rangle_{\mathcal{H}(k_P)} = \langle f, 0 \rangle_{\mathcal{H}(k_P)} = 0 \).
The important point here is that $D_{k_P}(Q)$ does not require integrals with respect to $P$ to be computed. A kernel $k_P$ with $\int k_P(\cdot, x)\,dP = 0$ will be called a *Stein kernel* (for $P$), for reasons that will become clear in the sequel. An example for how such a kernel can be constructed is as follows: Consider the bounded linear operator $(A_P g)(x) = g(x) - \int g\,dP$ acting on elements of an RKHS $\mathcal{H}(k)$. If we apply $A_P$ to both arguments of the kernel $k_P$, we obtain a Stein kernel $k_P(x, y) = A^y_P A^x_P k(x, y)$

$$= k(x, y) - \int k(x, y)\,dP(x) - \int k(x, y)\,dP(y) + \int \int k(x, y)\,dP(x)\,dP(y). \quad (10)$$

Indeed, $\int k_P(\cdot, x)\,dP(x) = \int A^y_P A^x_P k(x, y)\,dP(x) = A^y_P \int A^x_P k(x, y)\,dP(x) = A^y_P 0 = 0$, where interchange of $A^y_P$ and the integral is justified by noting that $A^y_P$ is a bounded linear operator and following similar reasoning to Lemma 2. In fact, the RKHS $\mathcal{H}(k_P)$ consists of functions of the form $A_P g = g - \int g\,dP$ where $g \in \mathcal{H}(k)$. Unfortunately, the Stein kernel in (10) is not useful because it still involves the problematic integral $\int k(\cdot, x)\,dP(x)$. The next section presents a more useful construction of a Stein kernel.

### 4.1 Stein Operators

The aim here is to identify an alternative operator $A_P$, which can be computed. Let $\nabla f = (\partial_{x_1} f, \ldots, \partial_{x_d} f)^\top$ for differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$. Our main tool is a Stein operator, and the classical example of this is as follows:

**Standing Assumption 3.** The distribution $P$ admits a positive and differentiable PDF such that $x \mapsto (\nabla \log p)(x)$ is Lipschitz.

**Definition 10** (Canonical Stein operator). For a distribution $P$ admitting a positive and differentiable density $p$ on $\mathbb{R}^d$, we define the canonical Stein operator

$$(A_P g)(x) = (\nabla \cdot g)(x) + g(x) \cdot (\nabla \log p)(x)$$

acting on differentiable vector field $g : \mathbb{R}^d \to \mathbb{R}^d$, where $x \in \mathbb{R}^d$.

The canonical Stein operator was introduced (for Gaussian $P$) in [Stein 1972]. Importantly, observe that

$$(\nabla \log p)(x) = \frac{(\nabla p)(x)}{p(x)} = \frac{1}{Z} (\nabla \tilde{p})(x) = \frac{(\nabla \tilde{p})(x)}{\tilde{p}(x)} = (\nabla \log \tilde{p})(x),$$

which can be computed without knowledge of $p$ or $Z$, provided $\tilde{p}$ and $\nabla \tilde{p}$ can be evaluated. Loosely speaking, we can apply the Stein operator $A_P$ in Definition 10 to a standard kernel $k$ to obtain the following Stein kernel:
Lemma 6. Suppose that \( k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \) is a symmetric positive definite kernel with \((x, y) \mapsto \partial^{(\alpha, \beta)} k(x, y)\) being continuous and uniformly bounded for all \(|\alpha|, |\beta| \leq 1\). Suppose \( \int \| \nabla \log p(x) \| dP(x) < \infty \) and that \( \sup_{\|x\| \geq r} r^{d-1} p(x) \to 0 \) as \( r \to \infty \). Then

\[
k_P(x, y) = \nabla_x \cdot \nabla_y k(x, y) + \nabla_x \log p(x) \cdot \nabla_y k(x, y) + \nabla_y \log p(y) \cdot \nabla_x k(x, y)
+ (\nabla_x \log p(x)) \cdot (\nabla_y \log p(y)) k(x, y)
\]

is a symmetric positive definite kernel with \( \int k_P(x, y) dP(y) = 0 \) for all \( x \in \mathbb{R}^d \).

Proof. First notice that

\[
k_P(x, y) = \mathcal{A}_P^y \begin{bmatrix} \vdots \\ \nabla_{x_i} k(x, y) + k(x, y) \nabla_{x_i} \log p(x) \end{bmatrix} = \mathcal{A}_P^y g(y)
\]

where, under our assumptions, (a) \( y \mapsto g(y) \) is bounded, and (b) \( y \mapsto \nabla_y \cdot g(y) \) is integrable with respect to \( P \). Thus it suffices to show that \( \int \mathcal{A}_P g dP = 0 \) for all vector fields \( g \) for which (a) and (b) hold.

Let \( g \) be such a vector field, and let \( B_r = \{ x \in \mathbb{R}^d : \| x \| \leq r \} \) and \( S_r = \{ x \in \mathbb{R}^d : \| x \| = r \} \). The main idea is to apply the divergence theorem (i.e. integrate by parts):

\[
\int \mathcal{A}_P g dP = \int (\nabla \cdot g) + g \cdot (\nabla \log p) dP = \int \nabla \cdot (pg)(x) dP = \lim_{r \to \infty} \int_{B_r} \nabla \cdot (pg)(x) dP = \lim_{r \to \infty} \oint_{S_r} p(x)(g(x) \cdot n(x)) dS
\]

where \( n(x) \) is the outward unit normal to \( S_r \) at \( x \). (The regularity assumptions ensure that the integrals \( \int (\nabla \cdot g) dP \) and \( \int g \cdot (\nabla \log p) dP \) exist.) Now

\[
\int_{S_r} p(x)(g(x) \cdot n(x)) dS \leq \| g \|_{\infty} \sup_{\| x \| \geq r} p(x) \int_{S_r} dS \leq \| g \|_{\infty} \sup_{\| x \| \geq r} p(x) \frac{2\pi^{d/2}}{\Gamma(d/2)} \int_{S_r} dS \to 0 \text{ as } r \to \infty,
\]

where we have used the formula for the surface area of the radius \( r \) sphere in \( \mathbb{R}^d \).

\[\square\]

Definition 11 (Kernel Stein discrepancy). With \( k_P \) defined in Lemma 6 we call \( D_{k_P} \) in Lemma 5 a kernel Stein discrepancy (KSD).
Lemma 7. For $Q_n = \sum_{i=1}^{n} w_i \delta(x_i)$, we have the explicit form of KSD:

$$D_{k_P}(Q_n) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j k_P(x_i, x_j)}$$

Proof. Immediate from the closed form expression for MMD in Section 3.1, with $k_P$ in place of $k$ and using the fact that $\int k_P(x, y) dP(y) = 0$ for all $x \in \mathbb{R}^d$ from Lemma 6. \[\square\]

As with MMD, we can establish properties analogous to characteristicness and convergence control for KSD. Here we focus on the stronger property of convergence control:

Theorem 3. Let $P$ be distantly dissipative, meaning that $\liminf_{r \to \infty} \kappa(r) > 0$ where

$$\kappa(r) = \inf \left\{ -2 \frac{(\nabla \log p)(x) - (\nabla \log p)(y), x - y}{\|x - y\|^2} : \|x - y\| = r \right\}.$$  

Consider the kernel $k(x, y) = (\sigma^2 + \|x - y\|^2)^{-\beta}$ for some fixed $\sigma > 0$ and a fixed exponent $\beta \in (0, 1)$. Then $D_{k_P}(Q_n) \to 0$ implies $Q_n \Rightarrow P$.

Proof. Theorem 8 in [Gorham and Mackey 2017]. \[\square\]

Theorem 3 justifies attempting to minimise KSD from the point of view of quantisation, which we will discuss next.

4.2 Optimal Quantisation

The simplest use of KSD for quantisation is as follows:

Exercise 3 (Optimal quantisation with KSD). Consider $P = \mathcal{N}(0, I)$ and $k(x, y) = (\sigma^2 + \|x - y\|^2)^{-1/2}$ on $X = \mathbb{R}^d$. (You may wish to focus on $d = 1$ or $d = 2$.)

(a) Verify that $P$ and $k$ satisfy the conditions of Theorem 3.

(b) Calculate (analytically) the Stein kernel $k_P(x, y)$.

(c) For a fixed value of $n$ (e.g. $n = 10$) and a fixed value of $\sigma$ (e.g. $\sigma = 1$), try to numerically optimise the locations of the states $x_1, \ldots, x_n$ in order to minimise $D_k(P, \frac{1}{n} \sum_{i=1}^{n} \delta(x_i))$.

(d) What effect does varying the bandwidth parameter $\sigma$ have on the approximations that are produced?

(e) What happens if instead the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/\sigma^2)$ is used?

This provides an optimisation-based alternative to popular sampling-based algorithms, such as MCMC and SMC. See Figure 8.
Stein Points

Figure 8: *Stein points* (left) and *Stein thinning* (right): Stein points are generated by sequential (greedy) minimisation of KSD to select $n = 8$ states $\{x_i\}_{i=1}^n$ in an approximation $Q_n$ to $P$. The numbers indicate the order in which the states $x_i$ were selected. Stein thinning restricts the continuous inner-loop optimisation problem in Stein points to a discrete search over a MCMC sample path (black).

Remark 14. The Gaussian distribution $P$ in Exercise 3 is analytically tractable, so KSD is not actually needed and MMD can be used. You are encouraged to extend Exercise 3 by replacing $P$ with an intractable posterior distribution arising in an application of Bayesian statistics.

Remark 15. Sequential greedy optimisation provides one practical solution to part (c) of Exercise 3, and was studied in detail in Chen et al. [2018, 2019]. States $\{x_i\}_{i=1}^n$ constructed in this way were called Stein points.

4.3 Optimal Approximation

In challenging applications of Bayesian statistics, the optimisation over $\mathcal{X}$ that was required to perform Exercise 3 will be difficult. Nevertheless, approximate sampling from $P$ may still be possible using MCMC or SMC. Through the optimisation of weights, Stein discrepancy provides an means to improve the approximations produced by MCMC or SMC, in a similar spirit to how importance sampling is sometimes used.

However, if we were to apply Lemma 4 with the kernel $k_P$ in place of $k$ we would obtain a degenerate solution, since $z_i = \int k_P(\cdot, x_i) dP = 0$ and optimal weights are $w = 0$. This makes sense, since we know that all $f \in \mathcal{H}(k_P)$ integrate to 0. In order to make progress we need an additional constraint on the weights, and for this purpose it is natural to impose that $w_1 + \cdots + w_n = 1$. This leads to the following extension of Lemma 4 which we present for a Stein kernel:
Lemma 8. Let $x_1, \ldots, x_n \in \mathcal{X}$ be distinct. The optimal weights

$$\arg \min_{w \in \mathbb{R}^n} D_{k_P} \left( \sum_{i=1}^{n} w_i \delta(x_i) \right)$$

are

$$w = \frac{K_P^{-1} \mathbf{1}}{1^\top K_P^{-1} \mathbf{1}}$$

where $[K_P]_{ij} = k_P(x_i, x_j)$.

Proof. From Lemma 7 we have

$$D_{k_P} \left( \sum_{i=1}^{n} w_i \delta(x_i) \right)^2 = w^\top K_P w,$$

so the optimisation problem is

$$\arg \min w^\top K_P w \quad \text{s.t.} \quad 1^\top w = 1.$$

This can be solved using the method of Lagrange multipliers to obtain the stated result. \qed

See the left panel of Figure 9. As with optimal weights for MMD, the linear system which must be solved can be numerically ill-conditioned when $n$ is large, or when two of the $x_i$ are very close or identical. Techniques for numerical regularisation could be used.

Exercise 4 (Bias correction with KSD). Consider again the distribution $P$ and kernel $k$ from Exercise 3.

(a) Generate and store a sequence $(x_i)_{i=1}^{100}$ of independent samples from $P$.

(b) For each $n = 1, \ldots, 100$, calculate and plot the values of

$$D_k \left( P, \frac{1}{n} \sum_{i=1}^{n} \delta(x_i) \right) \quad \text{and} \quad D_k \left( P, \sum_{i=1}^{n} w_i^{(n)} \delta(x_i) \right)$$

where $w^{(m)} = (w_1^{(m)}, \ldots, w_n^{(m)})$ are the optimal weights from Lemma 8 based on the states $\{x_i\}_{i=1}^{m}$. What do you observe?

(c) What happens if, instead, we consider $n = 1, \ldots, 1000$?

Remarkably, the use of Stein discrepancy in Exercise 4 can provide consistent approximations of $P$ even if the states $x_i$, generated in step (a), are not sampled from $P$ (provided, at least, that they are sampled from a distribution that is not too different from $P$). See Liu and Lee [2017], Hodgkinson et al. [2020] and Theorem 3 of Riabiz et al. [2021]. This is reminiscent of importance sampling, and Liu and Lee [2017] termed it black-box importance sampling.
In black box importance sampling the weights $w_1, \ldots, w_n$ are obtained by minimising KSD in the manner of Lemma 8. Blue indicates states $x_i$ with positive weights $w_i > 0$, while red indicates negative weights $w_i < 0$. The size of the circles is proportional to $|w_i|$. Stein variational gradient descent [Liu and Wang, 2016] is one of a number of recently developed algorithms based on KSD; here an initial discrete distribution (light blue circles) are evolved in time (black lines) toward an optimal quantisation (dark blue circles) of $P$.

**Exercise 5** (Pathologies of KSD). The purpose of this final exercise is to illustrate one of the main weaknesses of KSD; insensitivity to distant high probability regions. Consider the Gaussian mixture model

$$p(x) \propto \exp(-x^2) + \exp(-(x-c)^2)$$

for $c \geq 0$.

(a) Calculate (analytically) the gradient $(\nabla \log p)(x)$.

(b) Derive the kernel $k_P(x, y)$, using $k(x, y) = (1 + (x - y)^2)^{-1/2}$ (to ensure convergence control; c.f. Theorem 3).

(c) Generate and store an independent sample $(x_i)_{i=1}^n$ from $\mathcal{N}(0,1)$, with $n = 100$.

(d) Plot $D_{k_P}\left(\frac{1}{n} \sum_{i=1}^n \delta(x_i)\right)$ as a function of $c \in (0,10)$.

(e) What do you conclude about the sensitivity of KSD to distant high probability regions?

**Chapter Notes** The canonical Stein operator is sometimes called the *Langevin–Stein operator* due to its close connection with the generator of a Langevin diffusion process [Barbour, 1988, 1990, Gorham and Mackey, 2015]. The general concept of a Stein discrepancy, in Lemma 5, was introduced in Gorham and Mackey [2015]. The Stein kernel was introduced
in Oates et al. [2017] and KSD was later introduced simultaneously in Chwialkowski et al. [2016], Liu et al. [2016]. Lemma 6 can be found in South et al. [2021]. Theorem 3 was slightly generalised to allow for invertible linear transformations of $x$ in the kernel $k$ in Theorem 4 of Chen et al. [2019]. The optimal weights in Lemma 8 are numerically ill-conditioned when $n$ is large; to address this, Riabiz et al. [2021] showed that greedy subset selection can be almost as accurate, but with lower computational complexity. Stein discrepancy is an active research topic at the moment, with many extensions attracting attention, such as to non-Euclidean domains [Barp et al., 2021], to other function classes besides kernels [Si et al., 2020, Grathwohl et al., 2020], and inspiring new algorithms such as Stein variational gradient descent [Liu and Wang, 2016] (see the right panel of Figure 9). For a recent literature review, see Anastasiou et al. [2021].

References

Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, Lester Mackey, Chris J Oates, Gesine Reinert, and Yvik Swan. Stein’s method meets statistics: A review of some recent developments. arXiv:2105.03481, 2021.

Andrew D Barbour. Stein’s method and Poisson process convergence. Journal of Applied Probability, 25(A):175–184, 1988.

Andrew D Barbour. Stein’s method for diffusion approximations. Probability Theory and Related Fields, 84(3):297–322, 1990.

Alessandro Barp, Chris J Oates, Emilio Porcu, and Mark Girolami. A Riemann–Stein kernel method. Bernoulli, 2021. To appear.

Alain Berlinet and Christine Thomas-Agnan. Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer Science & Business Media, 2011.

Wilson Ye Chen, Lester Mackey, Jackson Gorham, François-Xavier Briol, and Chris J Oates. Stein points. In International Conference on Machine Learning, pages 844–853. PMLR, 2018.

Wilson Ye Chen, Alessandro Barp, François-Xavier Briol, Jackson Gorham, Mark Girolami, Lester Mackey, and Chris J Oates. Stein point Markov chain Monte Carlo. In International Conference on Machine Learning, pages 1011–1021. PMLR, 2019.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In International Conference on Machine Learning, pages 2606–2615. PMLR, 2016.

Josef Dick and Friedrich Pillichshammer. Digital nets and sequences: Discrepancy theory and quasi-Monte Carlo integration. Cambridge University Press, 2010.

Martin Ehler, Manuel Gräf, and Chris J Oates. Optimal Monte Carlo integration on closed manifolds. Statistics and Computing, 29(6):1203–1214, 2019.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein’s method. Advances in Neural Information Processing Systems, 28:226–234, 2015.
Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In International Conference on Machine Learning, pages 1292–1301. PMLR, 2017.

Manuel Gräf, Daniel Potts, and Gabriele Steidl. Quadrature errors, discrepancies, and their relations to halftoning on the torus and the sphere. *SIAM Journal on Scientific Computing*, 34(5):A2760–A2791, 2012.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Learning the Stein discrepancy for training and evaluating energy-based models without sampling. In International Conference on Machine Learning, pages 3732–3747. PMLR, 2020.

Fred Hickernell. A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, 67(221):299–322, 1998.

Liam Hodgkinson, Robert Salomone, and Fred Roosta. The reproducing Stein kernel approach for post-hoc corrected sampling. arXiv:2001.09266, 2020.

Toni Karvonen and Simo Särkkä. Fully symmetric kernel quadrature. *SIAM Journal on Scientific Computing*, 40(2):A697–A720, 2018.

Toni Karvonen, Chris J Oates, and Simo Särkkä. A Bayes–Sard cubature method. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 5886–5897, 2018.

Toni Karvonen, Chris J Oates, and Mark Girolami. Integration in reproducing kernel Hilbert spaces of Gaussian kernels. *Mathematics of Computation*, 90(331):2209–2233, 2021.

Lauwerens Kuipers and Harald Niederreiter. *Uniform Distribution of Sequences*. Wiley, 1974.

Gunther Leobacher and Friedrich Pillichshammer. *Introduction to quasi-Monte Carlo integration and applications*. Springer, 2014.

Qiang Liu and Jason Lee. Black-box importance sampling. In *Artificial Intelligence and Statistics*, pages 952–961. PMLR, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284. PMLR, 2016.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. arXiv:1605.09522, 2016.

Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society, Series B*, 79:695–718, 2017.

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. 2013. URL https://statweb.stanford.edu/~owen/mc/.

Luc Pronzato and Anatoly Zhigljavsky. Bayesian quadrature, energy minimization, and space-filling design. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):959–1011, 2020.

A Reznikov and EB Saff. The covering radius of randomly distributed points on a manifold. *International Mathematics Research Notices*, 2016(19):6065–6094, 2016.
Marina Riabiz, Wilson Chen, Jon Cockayne, Pawel Swietach, Steven A Niederer, Lester Mackey, and Chris J Oates. Optimal thinning of MCMC output. *Journal of the Royal Statistical Society, Series B*, 2021. To appear.

Shijing Si, Chris J Oates, Andrew B Duncan, Lawrence Carin, François-Xavier Briol, et al. Scalable control variates for Monte Carlo methods via stochastic optimization. In *Proceedings of the 14th International Conference on Monte Carlo & Quasi-Monte Carlo Methods in Scientific Computing*, 2020. To appear.

Carl-Johann Simon-Gabriel, Alessandro Barp, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *arXiv:2006.09268*, 2020.

Leah F South, Toni Karvonen, Chris Nemeth, Mark Girolami, and Chris J Oates. Semi-exact control functionals from Sard’s method. *Biometrika*, 2021. To appear.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, pages 583–602. University of California Press, 1972.

Onur Teymur, Jackson Gorham, Marina Riabiz, and Chris J Oates. Optimal quantisation of probability measures using maximum mean discrepancy. In *International Conference on Artificial Intelligence and Statistics*, pages 1027–1035. PMLR, 2021.

Holger Wendland. Error estimates for interpolation by compactly supported radial basis functions of minimal degree. *Journal of Approximation Theory*, 93(2):258–272, 1998.

Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004.
5 Partial Solutions to Exercises

This section contains solutions to analytic parts of the exercises in the main text. For the numerical part of the exercises, Matlab solutions are provided.

5.1 Exercise 1

(a) In dimension $d$ we have the kernel mean embedding

\[ \mu_P(x) = \left( \frac{\sigma^2}{2 + \sigma^2} \right)^{d/2} \exp \left\{ -\frac{1}{(2 + \sigma^2)} \| x \|^2 \right\}. \]

5.2 Exercise 2

(a) The maximum value operator $z \mapsto z_+$ that appears in the kernels of Example 6 is irrelevant when restricting to $\mathcal{X} = [0, 1]$, since $1 - |x - y| \geq 0$ for all $x, y \in [0, 1]$. Thus we consider the following kernels on $\mathcal{X} = [0, 1]$

\[
\begin{align*}
    k(x, y) &= (r = |x - y|, x, y \in [0, 1]) \\
    (1 - r) &\quad (s = 1) \\
    (1 - r)^2(3r + 1) &\quad (s = 2) \\
    (1 - r)^3(8r^2 + 5r + 1) &\quad (s = 3)
\end{align*}
\]

Splitting $\int_0^1 k(x, y)dy$ into a sum of $\int_0^x k(x, y)dy$ and $\int_x^1 k(x, y)dy$, the integrals become straightforward and we can evaluate them analytically:

| order $s$ | $\mu_P(x)$ | $\int k(x, y)dx dy$ |
|----------|-------------|---------------------|
| $s = 1$  | $-x^2 + x + \frac{a}{3}$ | $\frac{3}{3}$ |
| $s = 2$  | $x^4 - 2x^3 + x + \frac{a}{2}$ | $\frac{4}{5}$ |
| $s = 3$  | $-2x^3 + 8x^7 - \frac{3}{4}x^6 + 7x^5 - \frac{7}{4}x^3 + x + \frac{1}{4}$ | $\frac{1}{5}$ |

5.3 Exercise 3

(b) Let $u(x) = (\nabla \log p)(x)$ and consider the kernel $k(x, y) = (1 + \|x - y\|^2)^{-1/2}$. For the case where $P = \mathcal{N}(0, I)$ we have $u(x) = -x$. Then we compute

\[
\begin{align*}
    \nabla_x k(x, y) &= -\frac{1}{(1 + \|x - y\|^2)^{3/2}}(x - y) \\
    \nabla_y k(x, y) &= \frac{1}{(1 + \|x - y\|^2)^{3/2}}(x - y) \\
    \nabla_x \cdot \nabla_y k(x, y) &= -\frac{3\|x - y\|^2}{(1 + \|x - y\|^2)^{3/2}} + \frac{d}{(1 + \|x - y\|^2)^{3/2}}
\end{align*}
\]

which leads to

\[
\begin{align*}
    k_P(x, y) := \nabla_x \cdot \nabla_y k(x, y) + [\nabla_x k(x, y)]^\top u(y) + [\nabla_y k(x, y)]^\top u(x) + k(x, y)[u(x)]^\top u(y)
    &= -\frac{3\|x - y\|^2}{(1 + \|x - y\|^2)^{5/2}} + 2\beta \left[ \frac{d + [u(x) - u(y)](x - y)^\top}{(1 + \|x - y\|^2)^{3/2}} \right] + \frac{u(x)^\top u(y)}{(1 + \|x - y\|^2)^{1/2}}
\end{align*}
\]
Theorem 5 of [Gorham and Mackey 2017] proves that the Stein kernel $k_P$ based on the Gaussian kernel $k$ provides weak convergence control when $P$ is distantly dissipative (and $\nabla \log p$ is Lipschitz, under Standing Assumption 3). However, for $d \geq 3$, Theorem 6 of [Gorham and Mackey 2017] proves that the corresponding KSD does not provide weak convergence control.

5.4 Exercise 5

(a) The required gradient is

$$(\nabla \log p)(x) = -2 \left[ \frac{x \exp(-x^2) + (x - c) \exp(-(x - c)^2)}{\exp(-x^2) + \exp(-(x - c)^2)} \right].$$