Biometrika (2020), **103**, 1, *pp.* 1–20
Printed in Great Britain Advance Access publication on 31 July 2018

Supplementary material for “Semi-Exact Control Functionals From Sard’s Method”

By L. F. South

School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD 4000, Australia.

l1.south@qut.edu.au

T. KARVONEN

The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK.

tkarvonen@turing.ac.uk

C. NEMETH

Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, U.K.

c.nemeth@lancaster.ac.uk

M. GIROLAMI

Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K.

mag92@eng.cam.ac.uk

and C. J. OATES

School of Mathematics, Statistics & Physics, Newcastle University, Newcastle NE1 7RU, U.K.

chris.oates@ncl.ac.uk

1. Proof of Lemma 1

Lemmas 1 and 2 are stylised versions of similar results that can be found in earlier work, such as Chwialkowski et al. (2016); Liu et al. (2016); Oates et al. (2017). Our presentation differs in that we provide a convenient explicit sufficient condition, on the tails of $\|\nabla g\|$ for Lemma 1, and on the tails of $\|\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{y}}^\top k(x,y)\|$ and $\|\nabla_x \Delta_y k(x,y)\|$ for Lemma 2, for their conclusions to hold.

Proof. The stated assumptions on the differentiability of p and g imply that the vector field $p(x)\nabla_x g(x)$ is continuously differentiable on $\mathbb{R}^d$. The divergence theorem can therefore be applied, over any compact set $D \subset \mathbb{R}^d$ with piecewise smooth boundary ∂D, to reveal that

$$\begin{aligned}\int_D (\mathcal{L}g)(x)p(x)\mathrm{d}x &= \int_D \{\Delta_x g(x) + \nabla_x g(x) \cdot \nabla_x \log p(x)\} p(x)\mathrm{d}x \\ &= \int_D \left[\frac{1}{p(x)} \nabla_x \cdot \{p(x)\nabla_x g(x)\}\right] p(x)\mathrm{d}x\end{aligned}$$

$$
\begin{aligned}
&= \int_D \nabla_x \cdot \{p(x)\nabla_x g(x)\}\mathrm{d}x \\
&= \oint_{\partial D} p(x)\nabla_x g(x) \cdot n(x)\sigma(\mathrm{d}x),
\end{aligned}
$$

where $n(x)$ is the unit normal vector at $x \in \partial D$ and $\sigma(\mathrm{d}x)$ is the surface element at $x \in \partial D$. Next, we let $D = D_R = \{x : \|x\| \leq R\}$ be the ball in $\mathbb{R}^d$ with radius R, so that ∂D_R is the sphere $S_R = \{x : \|x\| = R\}$. The assumption $\|\nabla_x g(x)\| \leq C\|x\|^{-\delta}p(x)^{-1}$ in the statement of the lemma allows us to establish the bound

$$
\begin{aligned}
\left|\oint_{S_R} p(x)\nabla_x g(x)\cdot n(x)\sigma(\mathrm{d}x)\right| \leq \oint_{S_R} \left|p(x)\nabla_x g(x)\cdot n(x)\right| \sigma(\mathrm{d}x) &\leq \oint_{S_R} p(x) \left\|\nabla_x g(x)\right\| \sigma(\mathrm{d}x) \\
&\leq \oint_{S_R} C \left\|x\right\|^{-\delta} \sigma(\mathrm{d}x) \\
&= CR^{-\delta} \oint_{S_R} \sigma(\mathrm{d}x) \\
&= CR^{-\delta} \frac{2\pi^{d/2}}{\Gamma(d/2)} R^{d-1},
\end{aligned}
$$

where in the first and second inequalities we used Jensen's inequality and Cauchy–Schwarz, respectively, and in the final equality we have made use of the surface area of S_R. The assumption that $\delta > d-1$ is then sufficient to obtain the result:

$$
\left|\int (\mathcal{L}g)(x)p(x)\mathrm{d}x\right| = \lim_{R\to\infty} \left|\oint_{S_R} p(x)\nabla_x g(x)\cdot n(x)\sigma(\mathrm{d}x)\right| \leq \lim_{R\to\infty} C\frac{2\pi^{d/2}}{\Gamma(d/2)}R^{d-1-\delta} = 0.
$$

This completes the argument. □

2. Differentiating the Kernel

This appendix provides explicit forms of (8) for kernels k that are radial. First we present a generic result in Lemma S1 before specialising to the cases of the rational quadratic (Section 2.1), Gaussian (Section 2.2) and Matérn (Section 2.3) kernels.

Lemma S1. *Consider a radial kernel k, meaning that k has the form*

$$
k(x,y) = \Psi(z), \qquad z = \|x-y\|^2,
$$

where the function $\Psi\colon [0,\infty) \to \mathbb{R}$ is four times differentiable and $x, y \in \mathbb{R}^d$. Then (8) *simplifies to*

$$
\begin{aligned}
k_0(x,y) = {} & 16z^2\Psi^{(4)}(z) + 16(2+d)z\Psi^{(3)}(z) + 4(2+d)d\Psi^{(2)}(z) \\
& + 4\{2z\Psi^{(3)}(z) + (2+d)\Psi^{(2)}(z)\}\{u(x)-u(y)\}^\top(x-y) \\
& - 4\Psi^{(2)}(z)u(x)^\top(x-y)(x-y)^\top u(y) - 2\Psi^{(1)}(z)u(x)^\top u(y),
\end{aligned} \tag{S1}
$$

where $u(x) = \nabla_x \log p(x)$.

Proof. The proof is direct and based on have the following applications of the chain rule:

$$
\begin{aligned}
\nabla_x k(x,y) &= 2\Psi^{(1)}(z)(x-y), \\
\nabla_y k(x,y) &= -2\Psi^{(1)}(z)(x-y),
\end{aligned}
$$

$$\begin{aligned}
\Delta_x k(x,y) &= 4z\Psi^{(2)}(z) + 2d\Psi^{(1)}(z), \\
\Delta_y k(x,y) &= 4z\Psi^{(2)}(z) + 2d\Psi^{(1)}(z), \\
\partial_{x_i}\partial_{y_j} k(x,y) &= -4\Psi^{(2)}(z)(x_i - y_i)(x_j - y_j) - 2\Psi^{(1)}(z)\delta_{ij}, \\
\nabla_x \Delta_y k(x,y) &= 8z\Psi^{(3)}(z)(x-y) + 4(2+d)\Psi^{(2)}(z)(x-y), \\
\nabla_y \Delta_x k(x,y) &= -8z\Psi^{(3)}(z)(x-y) - 4(2+d)\Psi^{(2)}(z)(x-y), \\
\Delta_x \Delta_y k(x,y) &= 16z^2\Psi^{(4)}(z) + 16(2+d)z\Psi^{(3)}(z) + 4(2+d)d\Psi^{(2)}(z).
\end{aligned}$$

Upon insertion of these formulae into (8), the desired result is obtained. □

Thus for kernels that are radial, it is sufficient to compute just the derivatives $\Psi^{(j)}$ of the radial part.

2.1. *Rational Quadratic Kernel*

The rational quadratic kernel,

$$\Psi(z) = (1 + \lambda^{-2} z)^{-1},$$

has derivatives $\Psi^{(j)}(z) = (-1)^j \lambda^{-2j} j! (1 + \lambda^{-2} z)^{-j-1}$ for $j \geq 1$.

2.2. *Gaussian Kernel*

For the Gaussian kernel we have $\Psi(z) = \exp(-z/\lambda^2)$. Consequently,

$$\Psi^{(j)}(z) = (-1)^j \lambda^{-2j} \exp(-z/\lambda^2),$$

for $j \geq 1$.

2.3. *Matérn Kernels*

For a Matérn kernel of smoothness $\nu > 0$ we have

$$\Psi(z) = b c^\nu z^{\nu/2} \, \mathrm{K}_\nu(c\sqrt{z}\,), \quad b = \frac{2^{1-\nu}}{\Gamma(\nu)}, \quad c = \frac{\sqrt{2\nu}}{\lambda},$$

where Γ the Gamma function and K_ν the modified Bessel function of the second kind of order ν. By the use of the formula $\partial_z K_\nu(z) = -\mathrm{K}_{\nu-1}(z) - \frac{\nu}{z}\mathrm{K}_\nu(z)$ we obtain

$$\Psi^{(j)}(z) = (-1)^j \frac{b c^{\nu+j}}{2^j} z^{(\nu-j)/2} \, \mathrm{K}_{\nu-j}(c\sqrt{z}\,),$$

for $j = 1, \ldots, 4$. In order to guarantee that the kernel is twice continously differentiable, so that k_0 in (7) is well-defined, we require that $\lceil \nu \rceil > 2$. As a Matérn kernel induces a reproducing kernel Hilbert space that is norm-equivalent to the standard Sobolev space of order $\nu + \frac{d}{2}$ (Fasshauer & Ye, 2011, Example 5.7), the condition $\lceil \nu \rceil > 2$ implies, by the Sobolev imbedding theorem (Adams & Fournier, 2003, Theorem 4.12), that the functions in $\mathcal{H}(k)$ are twice continuously differentiable. Notice that $\Psi^{(3)}(z)$ and $\Psi^{(4)}(z)$ may not be defined at $z = 0$, in which case the terms $16z^2\Psi^{(4)}(z)$, $16(2+d)z\Psi^{(3)}(z)$ and $8z\Psi^{(3)}(z)$ in (S1) must be interpreted as limits as $z \to 0$ from the right.

3. Properties of $\mathcal{H}(k_0)$

The purpose of this appendix is to establish basic properties of the reproducing kernel Hilbert space $\mathcal{H}(k_0)$ of the kernel k_0 in (7). For convenience, in this appendix we abbreviate $\mathcal{H}(k_0)$ to $\mathcal{H}$. In Lemma S2 we clarify the reproducing kernel Hilbert space structure of $\mathcal{H}$. Then in Lemma S3 we establish square integrability of the elements of $\mathcal{H}$ and in Lemma S4 we establish the local smoothness of the elements of $\mathcal{H}$.

To state these results we require several items of notation: The notation $C^s(\mathbb{R}^d)$ denotes the set of s-times continuously differentiable functions on $\mathbb{R}^d$; i.e. $\partial^\alpha f \in C^0(\mathbb{R}^d)$ for all $|\alpha| \leq s$ where $C^0(\mathbb{R}^d)$ denotes the set of continuous functions on $\mathbb{R}^d$. For two normed spaces V and W, let $V \hookrightarrow W$ denote that V is continuously embedded in W, meaning that $\|v\|_W \leq C\|v\|_V$ for all $v \in V$ and some constant $C \geq 0$. In particular, we write $V \simeq W$ if and only if V and W are equal as sets and both $V \hookrightarrow W$ and $W \hookrightarrow V$. Let $\mathcal{L}^2(p)$ denote the vector space of square integrable functions with respect to p and equip this with the norm $\|h\|_{\mathcal{L}^2(p)} = \{\int h(x)^2 p(x) \mathrm{d}x\}^{1/2}$. For $h : \mathbb{R}^d \to \mathbb{R}$ and $D \subset \mathbb{R}^d$ we let $h|_D : D \to \mathbb{R}$ denote the restriction of h to D.

First we clarify the reproducing kernel Hilbert space structure of $\mathcal{H}$:

Lemma S2 (Reproducing kernel Hilbert space structure of $\mathcal{H}$). *Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive-definite kernel such that the regularity assumptions of Lemma 2 are satisfied. Let $\mathcal{H}$ denote the normed space of real-valued functions on $\mathbb{R}^d$ with norm*

$$\|h\|_{\mathcal{H}} = \inf_{\substack{h=\mathcal{L}g \\ g \in \mathcal{H}(k)}} \|g\|_{\mathcal{H}(k)}.$$

Then $\mathcal{H}$ admits the structure of a reproducing kernel Hilbert space with kernel $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by $\kappa(x, y) = k_0(x, y)$. That is, $\mathcal{H} = \mathcal{H}(k_0)$. Moreover, for $D \neq \emptyset$, let $\mathcal{H}|_D$ denote the normed space of real-valued functions on D with norm

$$\|h'\|_{\mathcal{H}|_D} = \inf_{\substack{h|_D = h' \\ h \in \mathcal{H}}} \|h\|_{\mathcal{H}}.$$

Then $\mathcal{H}|_D$ is a reproducing kernel Hilbert space with kernel $\kappa|_D : D \times D \to \mathbb{R}$ given by $\kappa|_D(x, y) = k_0(x, y)$. That is, $\mathcal{H}|_D = \mathcal{H}(\kappa|_D)$.

Proof. The first statement is an immediate consequence of Theorem 5 in Section 4.1 of Berlinet & Thomas-Agnan (2011). The second statement is an immediate consequence of Theorem 6 in Section 4.2 of Berlinet & Thomas-Agnan (2011). □

Next we establish when the elements of $\mathcal{H}$ are square-integrable functions with respect to p.

Lemma S3 (Square integrability of $\mathcal{H}$). *Let κ be a radial kernel satisfying the pre-conditions of Lemma S1. If $u_i = \nabla_{x_i} \log p(\boldsymbol{x}) \in \mathcal{L}^2(p)$ for each $i = 1, \ldots, d$, then $\mathcal{H} \hookrightarrow \mathcal{L}^2(p)$.*

Proof. From the representer theorem and Cauchy–Schwarz we have

$$\int h(x)^2 p(x) \mathrm{d}x \;=\; \int \langle h, \kappa(\cdot, x) \rangle_{\mathcal{H}}^2 p(x) \mathrm{d}x \;\leq\; \|h\|_{\mathcal{H}}^2 \int \kappa(x, x) p(x) \mathrm{d}x. \tag{S2}$$

Now, in the special case $k(x, y) = \Psi(z)$, $z = \|x - y\|^2$, the conclusion of Lemma S1 gives that $\kappa(x, x) = 4(2 + d)d\Psi^{(2)}(0) - 2\Psi^{(1)}(0)\|u(x)\|^2$, from which it follows that

$$0 \leq \int \kappa(x, x) p(x) \mathrm{d}x \;=\; 4(2 + d)d\Psi^{(2)}(0) - 2\Psi^{(1)}(0) \int \|u(x)\|^2 p(x) \mathrm{d}x \;=\; C^2. \tag{S3}$$

The combination of (S2) and (S3) establishes that $\|h\|_{\mathcal{L}^2(p)} \leq C\|h\|_{\mathcal{H}}$, which is the claimed result. □

Finally we turn to the regularity of the elements of $\mathcal{H}$, as quantified by their smoothness over suitable bounded sets $D \subset \mathbb{R}^d$. In what follows we will let $\mathcal{H}(k)$ be a reproducing kernel Hilbert space of functions in $\mathcal{L}^2(\mathbb{R}^d)$, the space of square Lebesgue integrable functions on $\mathbb{R}^d$, such that the norms

$$\|h\|_{\mathcal{H}(k)} \simeq \|h\|_{W_2^r(\mathbb{R}^d)} = \left(\sum_{|\alpha| \leq r} \|\partial^\alpha h\|^2_{\mathcal{L}^2(\mathbb{R}^d)} \right)^{\frac{1}{2}}$$

are equivalent. The latter is recognized as the standard Sobolev norm; this space is denoted $W_2^r(\mathbb{R}^d)$. For example, the Matérn kernel in Section 2.3 corresponds to $\mathcal{H}(k)$ with $r = \nu + \frac{d}{2}$. The Sobolev embedding theorem implies that $W_2^r(\mathbb{R}^d) \subset C^0(\mathbb{R}^d)$ whenever $r > \frac{d}{2}$.

The following result establishes the smoothness of $\mathcal{H}$ in terms of the differentiability of its elements. If the smoothness of f is known then k should be selected so that the smoothness of $\mathcal{H}$ matches it.

Lemma S4 (Smoothness of $\mathcal{H}$). *Let $r, s \in \mathbb{N}$ be such that $r > s + 2 + \frac{d}{2}$. If $\mathcal{H}(k) \simeq W_2^r(\mathbb{R}^d)$ and $\log p \in C^{s+1}(\mathbb{R}^d)$, then, for any open and bounded set $D \subset \mathbb{R}^d$, we have $\mathcal{H}|_D \hookrightarrow W_2^s(D)$.*

Proof. Under our assumptions, the kernel $\kappa|_D : D \times D \to \mathbb{R}$ from Lemma S2 is s-times continuously differentiable in the sense of Definition 4.35 of Steinwart & Christmann (2008). It follows from Lemma 4.34 of Steinwart & Christmann (2008) that $\partial_x^\alpha \kappa|_D(\cdot, x) \in \mathcal{H}|_D$ for all $x \in D$ and $|\alpha| \leq s$. From the reproducing property in $\mathcal{H}|_D$ and the Cauchy–Schwarz inequality we have, for $|\alpha| \leq s$,

$$|\partial^\alpha f(x)| = \left|\langle f, \partial^\alpha \kappa|_D(\cdot, x)\rangle_{\mathcal{H}|_D}\right| \leq \|f\|_{\mathcal{H}|_D} \left\|\partial^\alpha \kappa|_D(\cdot, x)\right\|_{\mathcal{H}|_D} = \|f\|_{\mathcal{H}|_D} \left\{ \partial_x^\alpha \partial_y^\alpha \kappa|_D(x, y)|_{y=x} \right\}^{1/2}.$$

See also Corollary 4.36 of Steinwart & Christmann (2008). Thus it follows from the definition of $W_2^s(D)$ and the reproducing property that

$$\begin{aligned}\|f\|^2_{W_2^s(D)} = \sum_{|\alpha| \leq s} \|\partial^\alpha f\|^2_{L_2(D)} &\leq \|f\|^2_{\mathcal{H}|_D} \sum_{|\alpha| \leq s} \left\| x \mapsto \partial_x^\alpha \partial_y^\alpha \kappa|_D(x, y)|_{y=x} \right\|^2_{L^2(D)} \\ &= \|f\|^2_{\mathcal{H}|_D} \left\| x \mapsto \kappa|_D(x, x) \right\|^2_{W_2^s(D)}.\end{aligned}$$

Now, from the definition of κ and using the fact that k is symmetric, we have

$$\kappa(x, x) = \Delta_x \Delta_y k(x, y)|_{y=x} + 2u(x)^\top \nabla_x \Delta_y k(x, y)|_{y=x} + u(x)^\top \left\{ \nabla_x \nabla_y^\top k(x, y)|_{y=x} \right\} u(x).$$

Our assumption that $\mathcal{H}(k) \simeq W_2^r(\mathbb{R}^d)$ with $r > s + 2 + \frac{d}{2}$ implies that each of the functions $x \mapsto \Delta_x \Delta_y k(x, y)|_{y=x}$, $\nabla_x \Delta_y k(x, y)|_{y=x}$ and $\nabla_x \nabla_y^\top k(x, y)|_{y=x}$, are $C^s(\mathbb{R}^d)$. In addition, our assumption that $\log p \in C^{s+1}(\mathbb{R}^d)$ implies that $x \mapsto u(x) \in C^s(\mathbb{R}^d)$. Thus $x \mapsto \kappa(x, x)$ is $C^s(\mathbb{R}^d)$ and in particular the boundedness of D implies that $\|x \mapsto \kappa|_D(x, x)\|_{W_2^s(D)} < \infty$ as required. □

4. Proof of Lemma 2

Proof. In what follows C is a generic positive constant, independent of x but possibly dependant on y, whose value can differ each time it is instantiated. The aim of this proof is to apply Lemma 1 to the function $g(x) = \mathcal{L}_y k(x, y)$. Our task is to verify the pre-condition $\|\nabla_x g(x)\| \leq C\|x\|^{-\delta} p(x)^{-1}$ for some $\delta > d - 1$. It will then follow from the conclusion of Lemma 1 that $\int k_0(x, y) p(x) \, \mathrm{d}x = 0$ as required. To this end, expanding the term $\|\nabla_x g(x)\|^2$, we have

$$\begin{align*}
\|\nabla_x g(x)\|^2 &= \|\nabla_x \mathcal{L}_y k(x,y)\|^2 \\
&= \big\|\nabla_x \Delta_y k(x,y) + \nabla_x\{\nabla_y \log p(y) \cdot \nabla_y k(x,y)\}\big\|^2 \\
&= \|\nabla_x \Delta_y k(x,y)\|^2 + 2\nabla_x\big\{\nabla_y \log p(y) \cdot \nabla_y k(x,y)\big\}^\top \nabla_x \Delta_y k(x,y) \\
&\quad + \big\|\nabla_x\big\{\nabla_y \log p(y) \cdot \nabla_y k(x,y)\big\}\big\|^2 \\
&= \|\nabla_x \Delta_y k(x,y)\|^2 + 2\big[\{\nabla_x \nabla_y^\top k(x,y)\}^\top \nabla_y \log p(y)\big]^\top \nabla_x \Delta_y k(x,y) \\
&\quad + \big\|\{\nabla_x \nabla_y^\top k(x,y)\}\nabla_y \log p(y)\big\|^2 \\
&\leq \|\nabla_x \Delta_y k(x,y)\|^2 + 2\big\|\{\nabla_x \nabla_y^\top k(x,y)\}^\top \nabla_y \log p(y)\big\|\|\nabla_x \Delta_y k(x,y)\| \\
&\quad + \big\|\{\nabla_x \nabla_y^\top k(x,y)\}\nabla_y \log p(y)\big\|^2 \tag{S4} \\
&\leq \|\nabla_x \Delta_y k(x,y)\|^2 + 2\|\nabla_x \nabla_y^\top k(x,y)\|_{\mathrm{OP}}\|\nabla_y \log p(y)\|\|\nabla_x \Delta_y k(x,y)\| \\
&\quad + \|\nabla_x \nabla_y^\top k(x,y)\|_{\mathrm{OP}}^2 \|\nabla_y \log p(y)\|^2 \tag{S5} \\
&\leq \big\{C\|x\|^{-\delta} p(x)^{-1}\big\}^2 + 2\big\{C\|x\|^{-\delta} p(x)^{-1}\big\}\|\nabla_y \log p(y)\|\big\{C\|x\|^{-\delta} p(x)^{-1}\big\} \\
&\quad + \big\{C\|x\|^{-\delta} p(x)^{-1}\big\}^2 \|\nabla_y \log p(y)\|^2 \tag{S6} \\
&\leq C\|x\|^{-2\delta} p(x)^{-2}
\end{align*}$$

as required. Here (S4) follows from the Cauchy–Schwarz inequality applied to the second term, (S5) follows from the definition of the operator norm $\|\cdot\|_{\mathrm{OP}}$ and (S6) employs the pre-conditions that we have assumed. □

5. Proof of Lemma 3

Proof. Our first task is to establish that it is sufficient to prove the result in just the particular case $\hat{x}_N = 0$ and $N^{-1}I(\hat{x}_N)^{-1} = I$, where I is the d-dimensional identity matrix. Indeed, if $\hat{x}_N \neq 0$ or $N^{-1}I(\hat{x}_N)^{-1} \neq I$, then let $t(x) = W(x - \hat{x}_N)$ where W is a non-singular matrix satisfying $W^\top W = NI(\hat{x}_N)$ so that $t(x) \sim \mathcal{N}(0, I)$. Under the same co-ordinate transformation the polynomial subspace

$$A = \mathcal{P}_0^r = \mathrm{span}\{x^\alpha : \alpha \in \mathbb{N}_0^d,\ 0 \leq |\alpha| \leq r\}$$

becomes $B = \mathrm{span}\{t(x)^\alpha : \alpha \in \mathbb{N}_0^d,\ 0 \leq |\alpha| \leq r\}$. Exact integration of functions in A with respect to $\mathcal{N}(\hat{x}_N, I)$ corresponds to exact integration of functions in B with respect to $\mathcal{N}(0, I)$. Thus our first task is to establish that $B = A$. Clearly B is a linear subspace of A, since elements of B can be expanded out into monomials and monomials generate A, so it remains to argue that B is all of A. In what follows we will show that $\dim(B) = \dim(A)$ and this will complete the first part of the argument.

The co-ordinate transform t is an invertible affine map on $\mathbb{R}^d$. The action of such a map t on a set S of functions on $\mathbb{R}^d$ can be defined as $t(S) = \{x \to s(t(x)) : s \in S\}$. Thus $B = t(A)$. Let $t^*(x) = W^{-1}x + \hat{x}_n$ and notice that this is also an invertible affine map on $\mathbb{R}^d$ with $t^*(t(x)) = x$ being the identity map on $\mathbb{R}^d$. The composition of invertible affine maps on $\mathbb{R}^d$ is again an invertible affine map and thus t^*t is also an invertible affine map on $\mathbb{R}^d$ and its action on a set is well-defined. Considering the action of t^*t on the set A

gives that $t^*(t(A)) = A$ and therefore $t(A)$ must have the same dimension as A. Thus $\dim(A) = \dim(t(A)) = \dim(B)$ as claimed.

Our second task is to show that, in the case where p is the density of $\mathcal{N}(0, I)$ and thus $\nabla_x \log p(x) = -x$, the set $\mathcal{F} = \operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^r$ on which I_{CV} is exact is equal to $\mathcal{P}_0^r$. Our proof proceeds by induction on the maximal degree r of the polynomial. For the base case we take $r = 1$:

$$\begin{aligned}
\operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^1 &= \operatorname{span}\{1\} \oplus \operatorname{span}\big\{\mathcal{L}x_j : j = 1, \dots, d\big\} \\
&= \operatorname{span}\{1\} \oplus \operatorname{span}\big\{\Delta_x x_j + \nabla_x \log p(x) \cdot \nabla_x(x_j) : j = 1, \dots, d\big\} \\
&= \operatorname{span}\{1\} \oplus \operatorname{span}\big\{0 - x \cdot e_j : j = 1, \dots, d\big\} \\
&= \operatorname{span}\{1\} \oplus \operatorname{span}\big\{ - x_j : j = 1, \dots, d\big\} \\
&= \mathcal{P}_0^1.
\end{aligned}$$

For the inductive step we assume that $\operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^{r-1} = \mathcal{P}_0^{r-1}$ holds for a given $r \geq 2$ and aim to show that $\operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^r = \mathcal{P}_0^r$. Note that the action of $\mathcal{L}$ on a polynomial of order r will return a polynomial of order at most r, so that $\operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^r \subseteq \mathcal{P}_0^r$ and thus we need to show that $\mathcal{P}_0^r \subseteq \operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^r$. Under the inductive assumption we have

$$\begin{aligned}
\operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^r &= \operatorname{span}\{1\} \oplus \Big(\mathcal{L}\mathcal{P}^{r-1} \oplus \operatorname{span}\big\{\mathcal{L}x^\alpha :\, \alpha \in \mathbb{N}_0^d,\, |\alpha| = r\big\}\Big) \\
&= \Big(\operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^{r-1}\Big) \oplus \operatorname{span}\big\{\mathcal{L}x^\alpha :\, \alpha \in \mathbb{N}_0^d,\, |\alpha| = r\big\} \\
&= \mathcal{P}_0^{r-1} \oplus \operatorname{span}\big\{\mathcal{L}x^\alpha :\, \alpha \in \mathbb{N}_0^d,\, |\alpha| = r\big\} \\
&= \mathcal{P}_0^{r-1} \oplus \operatorname{span}\big\{\Delta_x x^\alpha + \nabla_x x^\alpha \cdot \nabla_x \log p(x) :\, \alpha \in \mathbb{N}_0^d,\, |\alpha| = r\big\} \\
&= \mathcal{P}_0^{r-1} \oplus \underbrace{\operatorname{span}\left\{ \sum_{j=1}^d \alpha_j(\alpha_j - 1) x_j^{\alpha_j - 2} \prod_{k \neq j} x_k^{\alpha_k} - \sum_{j=1}^d \alpha_j x^\alpha :\, \alpha \in \mathbb{N}_0^d,\, |\alpha| = r \right\}}_{=:\mathcal{Q}^r}
\end{aligned}$$

To complete the inductive step we must therefore show that, for each $\alpha \in \mathbb{N}_0^d$ with $|\alpha| = r$, we have $x^\alpha \in \operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^r$. Fix any $\alpha \in \mathbb{N}_0^d$ such that $|\alpha| = r$. Then

$$\phi(x) = \sum_{j=1}^d \alpha_j(\alpha_j - 1) x_j^{\alpha_j - 2} \prod_{k \neq j} x_k^{\alpha_k} - \sum_{j=1}^d \alpha_j x^\alpha \in \mathcal{Q}^r.$$

and

$$\varphi(x) = \frac{1}{1^\top \alpha} \sum_{j=1}^d \alpha_j(\alpha_j - 1) x_j^{\alpha_j - 2} \prod_{k \neq j} x_k^{\alpha_k} \in \mathcal{P}_0^{r-1}$$

because this polynomial is of order less than r. Since $\varphi - (1^\top \alpha)^{-1} \phi \in \mathcal{P}_0^{r-1} \oplus \mathcal{Q}^r = \operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^r$ and

$$\varphi(x) - \frac{1}{1^\top \alpha} \phi(x) = \frac{\sum_{j=1}^d \alpha_j}{1^\top \alpha} x^\alpha = x^\alpha,$$

we conclude that $x^\alpha \in \operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^r$. Thus we have shown that $\{x^\alpha : \alpha \in \mathbb{N}_0^d,\, |\alpha| = r\} \subset \operatorname{span}\{1\} \oplus \mathcal{L}\mathcal{P}^r$ and this completes the argument. □

6. Proof of Lemma 4

Proof. The assumptions that the $x^{(i)}$ are distinct and that k_0 is a positive-definite kernel imply that the matrix K_0 is positive-definite and thus non-singular. Likewise, the assumption that the $x^{(i)}$ are $\mathcal{F}$-unisolvent implies that the matrix P has full rank. It follows that the block matrix in (13) is non-singular. The interpolation and semi-exactness conditions in Section 2.3 can be written in matrix form as

1. $K_0 a + Pb = f$ (interpolation);
2. $P^\top a = 0$ (semi-exact).

The first of these is merely (10) in matrix form. To see how $P^\top a = 0$ is related to the semi-exactness requirement ($f_n = f$ whenever $f \in \mathcal{F}$), observe that for $f \in \mathcal{F}$ we have $f = Pc$ for some $c \in \mathbb{R}^m$. Consequently, the interpolation condition should yield $b = c$ and $a = 0$. The condition $P^\top a = 0$ enforces that $a = 0$ in this case: multiplication of the interpolation equation with $a^\top$ yields $a^\top K_0 a + a^\top Pb = a^\top Pc$, which is then equivalent to $a^\top K_0 a = 0$. Because K_0 is positive-definite, the only possible $a \in \mathbb{R}^n$ is $a = 0$ and P having full rank implies that $b = c$. Thus the coefficients a and b can be cast as the solution to the linear system

$$\begin{bmatrix} K_0 & P \\ P^\top & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}.$$

From (13) we get

$$b = (P^\top K_0^{-1} P)^{-1} P^\top K_0^{-1} f,$$

where $P^\top K_0^{-1} P$ is non-singular because K_0 is non-singular and P has full rank. Recognising that $b_1 = e_1^\top b$ for $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^m$ completes the argument. □

7. Nyström Approximation and Conjugate Gradient

In this appendix we describe how a Nyström approximation and the conjugate gradient method can be used to provide an approximation to the proposed method with reduced computational cost. To this end we consider a function of the form

$$\tilde{f}_{n_0}(x) = \tilde{b}_1 + \sum_{i=1}^{m-1} \tilde{b}_{i+1} \mathcal{L}\phi_i(x) + \sum_{i=1}^{n_0} \tilde{a}_i k_0(x, x^{(i)}), \tag{S7}$$

where $n_0 \ll n$ represents a small subset of the n points in the dataset. Strategies for selection of a suitable subset are numerous (e.g., Alaoui & Mahoney, 2015; Rudi et al., 2015) but for simplicity in this work a uniform random subset was selected. Without loss of generality we denote this subset by the first n_0 indices in the dataset. The coefficients a and b in the proposed method (10) can be characterized as the solution to a kernel least-squares problem, the details of which are reserved for Appendix 7.1. From this perspective it is natural to define the reduced coefficients $\tilde{a}$ and $\tilde{b}$ in (S7) also as the solution to a kernel least-squares problem, the details of which are reserved for Appendix 7.2. In taking this approach, the $(n+m)$-dimensional linear system in (13) becomes the (n_0+m)-dimensional linear system

$$\begin{bmatrix} K_{0,n_0,n} K_{0,n,n_0} + P_{n_0} P_{n_0}^\top & K_{0,n_0,n} P \\ P^\top K_{0,n,n_0} & P^\top P \end{bmatrix} \begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} = \begin{bmatrix} K_{0,n_0,n} f \\ P^\top f \end{bmatrix}. \tag{S8}$$

Here $K_{0,r,s}$ denotes the matrix formed by the first r rows and the first s columns of K_0. Similarly P_r denotes the first r rows of P. It can be verified that there is no approximation error when $n_0 = n$, with $\tilde{a} = a$ and $\tilde{b} = b$. This is a simple instance of a Nyström approximation and it can be viewed as a random projection method (Smola & Schökopf, 2000; Williams & Seeger, 2001).

The computational complexity of computing this approximation to the proposed method is

$$O(nn_0^2 + nm^2 + n_0^3 + m^3),$$

which could still be quite high. For this reason, we now consider iterative, as opposed to direct, linear solvers for (S8). In particular, we employ the conjugate gradient method to approximately solve this linear system. The performance of the conjugate gradient method is determined by the condition number of the linear system, and for this reason a preconditioner should be employed[1]. In this work we considered the preconditioner

$$\begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}.$$

Following Rudi et al. (2017), B_1 is the lower-triangular matrix resulting from a Cholesky decomposition

$$B_1 B_1^\top = \left(\frac{n}{n_0} K_{0,n_0,n_0}^2 + P_{n_0} P_{n_0}^\top \right)^{-1},$$

the latter being an approximation to the inverse of $K_{0,n_0,n} K_{0,n,n_0} + P_{n_0} P_{n_0}^\top$ and obtained at $O(n_0^3 + mn_0^2)$ cost. The matrix B_2 is

$$B_2 B_2^\top = \left(P^\top P\right)^{-1},$$

which uses the pre-computed matrix $P^\top P$ and is of $O(m^3)$ complexity. Thus we obtain a preconditioned linear system

$$\begin{bmatrix} B_1^\top (K_{0,n_0,n} K_{0,n,n_0} + P_{n_0} P_{n_0}^\top) B_1 & B_1^\top K_{0,n_0,n} P B_2 \\ B_2^\top P^\top K_{0,n,n_0} B_1 & I \end{bmatrix} \begin{bmatrix} \tilde{\tilde{a}} \\ \tilde{\tilde{b}} \end{bmatrix} = \begin{bmatrix} B_1^\top K_{0,n_0,n} f \\ B_2^\top P^\top f \end{bmatrix}.$$

The coefficients $\tilde{a}$ and $\tilde{b}$ of $\tilde{f}_{n_0}$ are related to the solution $(\tilde{\tilde{a}}, \tilde{\tilde{b}})$ of this preconditioner linear system via $\tilde{\tilde{a}} = B_1^{-1} \tilde{a}$ and $\tilde{\tilde{b}} = B_2^{-1} \tilde{b}$, which is an upper-triangular linear system solved at quadratic cost.

The above procedure leads to a more computationally (time and space) efficient procedure, and we denote the resulting estimator as $I_{\text{ASECF}}(f) = \tilde{b}_1$. Further extensions could be considered; for example non-uniform sampling for the random projection via leverage scores (Rudi et al., 2015).

For the examples in Section 3, we consider $n_0 = \lceil \sqrt{n} \, \rceil$ where $\lceil \cdot \rceil$ denotes the ceiling function. We use the `R` package `Rlinsolve` to perform conjugate gradient, where we specify the tolerance to be 10^{-5}. The initial value for the conjugate gradient procedure was the choice of $\tilde{\tilde{a}}$ and $\tilde{\tilde{b}}$ that leads to the Monte Carlo estimate, $\tilde{\tilde{a}} = 0$ and $\tilde{\tilde{b}} = B_2^{-1} e_1 \frac{1}{n} \sum_{i=1}^{n} f(x^{(i)})$. In our examples, we did not see a computational speed up from the use of conjugate gradient, likely due to the relatively small values of n involved.

[1] A linear system $Ax = b$ can be *preconditioned* by an invertible matrix P to produce $P^\top APz = P^\top b$. The solution z is related to x via $x = Pz$.

7.1. Kernel Least-Squares Characterization

Here we explain how the interpolant f_n in (10) can be characterized as the solution to the constrained kernel least-squares problem

$$\arg\min_{a,b} \frac{1}{n}\sum_{i=1}^{n}\left[f(x^{(i)}) - f_n(x^{(i)})\right]^2 \quad \text{s.t.} \quad f_n = f \quad \text{for all} \quad f \in \mathcal{F}.$$

To see this, note that similar reasoning to that in Appendix 6 allows us to formulate the problem using matrices as

$$\arg\min_{a,b} \|f - K_0 a - Pb\|^2 \quad \text{s.t.} \quad P^\top a = 0. \tag{S9}$$

This is a quadratic minimization problem subject to the constraint $P^\top a = 0$ and therefore the solution is given by the Karush–Kuhn–Tucker matrix equation

$$\begin{bmatrix} K_0^2 & K_0 P & P \\ P^\top K_0 & P^\top P & 0 \\ P^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} Kf \\ P^\top f \\ 0 \end{bmatrix}. \tag{S10}$$

Now, we are free to add a multiple, P, of the third row to the first row, which produces

$$\begin{bmatrix} K_0^2 + PP^\top & K_0 P & P \\ P^\top K_0 & P^\top P & 0 \\ P^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} Kf \\ P^\top f \\ 0 \end{bmatrix}.$$

Next, we make the *ansatz* that $c = 0$ and seek a solution to the reduced linear system

$$\begin{bmatrix} K_0^2 + PP^\top & K_0 P \\ P^\top K_0 & P^\top P \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} Kf \\ P^\top f \end{bmatrix}.$$

This is the same as

$$\begin{bmatrix} K_0 & P \\ P^\top & 0 \end{bmatrix} \begin{bmatrix} K_0 & P \\ P^\top & 0 \end{bmatrix} = \begin{bmatrix} K_0 & P \\ P^\top & 0 \end{bmatrix} \begin{bmatrix} f \\ 0 \end{bmatrix}$$

and thus, if the block matrix can be inverted, we have

$$\begin{bmatrix} K_0 & P \\ P^\top & 0 \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix} \tag{S11}$$

as claimed. Existence of a solution to (S11) establishes a solution to the original system (S10) and justifies the *ansatz.* Moreover, the fact that a solution to (S11) exists was established in Lemma 4.

7.2. Nyström Approximation

To develop a Nyström approximation, our starting point is the kernel least-squares characterization of the proposed estimator in (S9). In particular, the same least-squares problem can be considered for the Nyström approximation in (S7):

$$\arg\min_{\tilde{a},\tilde{b}} \|f - K_{0,n,n_0}\tilde{a} - P\tilde{b}\|_2^2 \quad \text{s.t.} \quad P_{n_0}^\top \tilde{a} = 0.$$

This least-squares problem can be formulated as

$$\begin{aligned}
&\arg\min_{\tilde{a},\tilde{b}}(f - K_{0,n,n_0}\tilde{a} - P\tilde{b})^\top (f - K_{0,n,n_0}\tilde{a} - P\tilde{b}) \\
&= \arg\min_{\tilde{a},\tilde{b}} \Big(f^\top f - f^\top K_{0,n,n_0}\tilde{a} - f^\top P\tilde{b} - \tilde{a}^\top K_{0,n_0,n} f + \tilde{a}^\top K_{0,n_0,n} K_{0,n,n_0}\tilde{a} \\
&\qquad + \tilde{a}^\top K_{0,n_0,n} P\tilde{b} - \tilde{b}^\top P^\top f - \tilde{b}^\top P^\top K_{0,n,n_0}\tilde{a} - \tilde{b}^\top P^\top P \tilde{b} \Big) \\
&= \arg\min_{\tilde{a},\tilde{b}} \begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix}^\top \begin{bmatrix} K_{0,n_0,n}K_{0,n,n_0} & K_{0,n_0,n}P \\ P^\top K_{0,n,n_0} & P^\top P \end{bmatrix} \begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} - 2 \begin{bmatrix} K_{0,n_0,n} f \\ P^\top f \end{bmatrix} \begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} + f^\top f
\end{aligned}$$

This is a quadratic minimization problem subject to the constraint $P_{n_0}^\top \tilde{a} = 0$ and so the solution is given by the Karush–Kuhn–Tucker matrix equation

$$\begin{bmatrix} K_{0,n_0,n}K_{0,n,n_0} & K_{0,n_0,n}P & P_{n_0} \\ P^\top K_{0,n,n_0} & P^\top P & 0 \\ P_{n_0}^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{a} \\ \tilde{b} \\ \tilde{c} \end{bmatrix} = \begin{bmatrix} K_{0,n_0,n} f \\ P^\top f \\ 0 \end{bmatrix}. \tag{S12}$$

Following an identical argument to that in Appendix 7.1, we first add P_{n_0} times the third row to the first row to obtain

$$\begin{bmatrix} K_{0,n_0,n}K_{0,n,n_0} + P_{n_0}P_{n_0}^\top & K_{0,n_0,n}P & P_{n_0} \\ P^\top K_{0,n,n_0} & P^\top P & 0 \\ P_{n_0}^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{a} \\ \tilde{b} \\ \tilde{c} \end{bmatrix} = \begin{bmatrix} K_{0,n_0,n} f \\ P^\top f \\ 0 \end{bmatrix}.$$

Taking again the *ansatz* that $\tilde{c} = 0$ requires us to solve the reduced linear system

$$\begin{bmatrix} K_{0,n_0,n}K_{0,n,n_0} + P_{n_0}P_{n_0}^\top & K_{0,n_0,n}P \\ P^\top K_{0,n,n_0} & P^\top P \end{bmatrix} \begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} = \begin{bmatrix} K_{0,n_0,n} f \\ P^\top f \end{bmatrix}. \tag{S13}$$

As in Appendix 7.1, the existence of a solution to (S13) implies a solution to (S12) and justifies the *ansatz*.

8. Sensitivity to the Choice of Kernel

In this appendix we investigate the sensitivity of kernel-based methods (I_{CF}, I_{SECF} and I_{ASECF}) to the kernel and its parameter using the Gaussian example of Section 3.2. Specifically we compare the three kernels described in Appendix 2, the Gaussian, Matérn and rational quadratic kernels, when the parameter, λ, is chosen using either cross-validation or the median heuristic (Garreau et al., 2017). For the Matérn kernel, we fix the smoothness parameter at $\nu = 4.5$.

In the cross-validation approach,

$$\lambda_{\mathrm{CV}} \in \arg\min \sum_{i=1}^{5} \sum_{j=1}^{n_5} \left\{ f(x^{(i,j)}) - f_{i,\lambda}(x^{(i,j)}) \right\}^2, \tag{S14}$$

where $n_5 := \lfloor n/5 \rfloor$, $f_{i,\lambda}$ denotes an interpolant of the form (10) to f at the points $\{x^{(i,j)} : j = 1, \dots, n_5\}$ with kernel parameter λ, and $x^{(i,j)}$ is the jth point in the ith fold. In general (S14) is an intractable optimization problem and we therefore perform a grid-based search. Here we consider $\lambda \in 10^{\{-1.5,-1,-0.5,0,0.5,1\}}$. 305

The median heuristic described in Garreau et al. (2017) is the choice of the bandwidth

$$\tilde{\lambda} = \sqrt{\frac{1}{2}\text{Med}\Big\{\|x^{(i)} - x^{(j)}\|^2 \;:\; 1 \leq i < j \leq n\Big\}}$$

for functions of the form $k(x, y) = \varphi(\|x - y\|/\lambda)$, where Med is the empirical median. This heuristic can be used for the Gaussian, Matérn and rational quadratic kernels, which all fit into this framework.

310 Figures S1 and S2 show the statistical efficiency of each combination of kernel and tuning approach for $n = 1000$ and $d = 4$, respectively. The outcome that the performance of I_{SECF} and I_{ASECF} are less sensitive to the kernel choice than I_{CF} is intuitive when considering the fact that semi-exact control functionals enforce exactness on $f \in \mathcal{F}$.

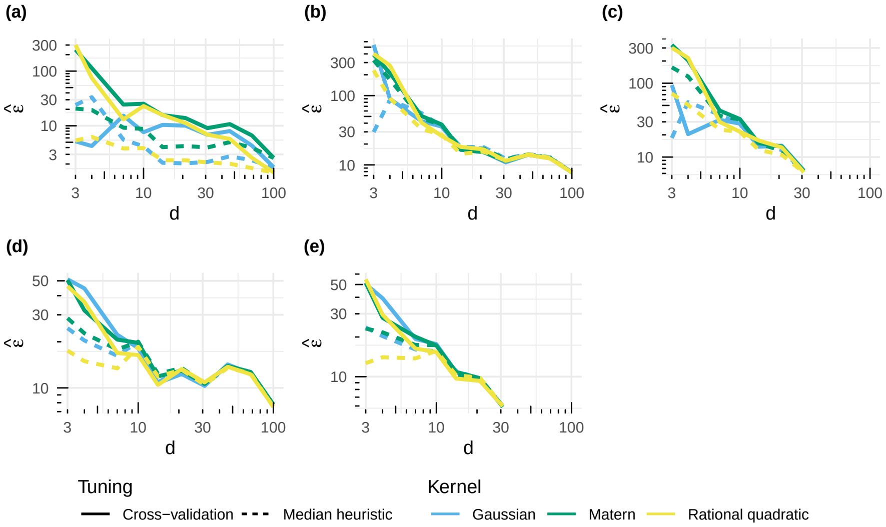

Fig. S1: Gaussian example, estimated statistical efficiency for $n = 1000$ using different kernels and tuning approaches. The estimators are (a) I_{CF}, (b) I_{SECF} with polynomial order $r = 1$, (c) I_{SECF} with $r = 2$, (d) I_{ASECF} with $r = 1$ and (e) I_{ASECF} with $r = 2$.

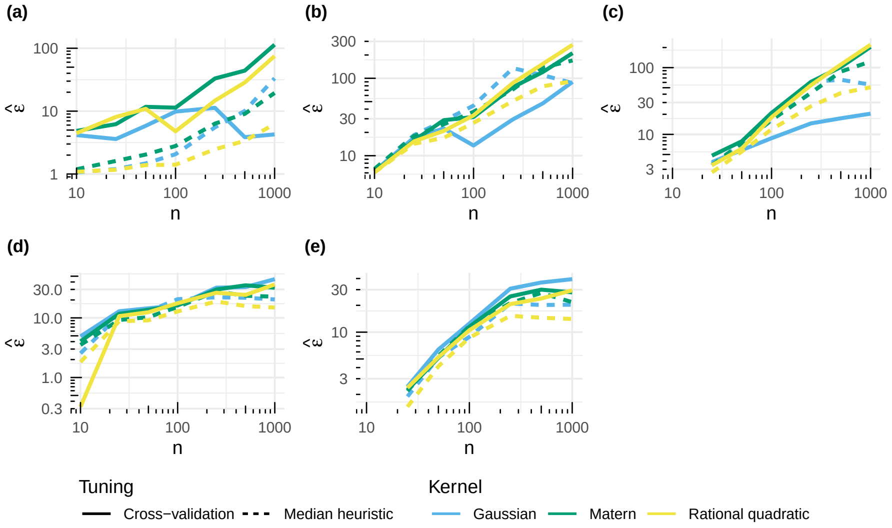

Fig. S2: Gaussian example, estimated statistical efficiency for $d = 4$ using different kernels and tuning approaches. The estimators are (a) I_{CF}, (b) I_{SECF} with polynomial order $r = 1$, (c) I_{SECF} with $r = 2$, (d) I_{ASECF} with $r = 1$ and (e) I_{ASECF} with $r = 2$.

9. Results for the Unadjusted Langevin Algorithm

Recall that the proposed method does not require that the $x^{(i)}$ form an empirical approximation to p. It is therefore interesting to investigate the behaviour of the method when the $(x^{(i)})_{i=1}^{\infty}$ arise as a Markov chain that does not leave p invariant. Figures S3 and S4 show results when the unadjusted Langevin algorithm is used rather than the Metropolis-adjusted Langevin algorithm which is behind Figures 3 and 4 of the main text. The benefit of the proposed method for samplers that do not leave p invariant is evident through its reduced bias compared to I_{ZV} and I_{MC} in Figure S5. Recall that the unadjusted Langevin algorithm (Parisi, 1981; Ermak, 1975) is defined by

$$x^{(i+1)} = x^{(i)} + \frac{h^2}{2}\Sigma\nabla_x \log P_{x|y}(x^{(i)} \mid y) + \epsilon_{i+1},$$

for $i = 1, \ldots, n-1$ where $x^{(1)}$ is a fixed point with high posterior support and $\epsilon_{i+1} \sim \mathcal{N}(0, h^2\Sigma)$. Step sizes of $h = 0.9$ for the sonar example and $h = 1.1$ for the capture-recapture example were selected.

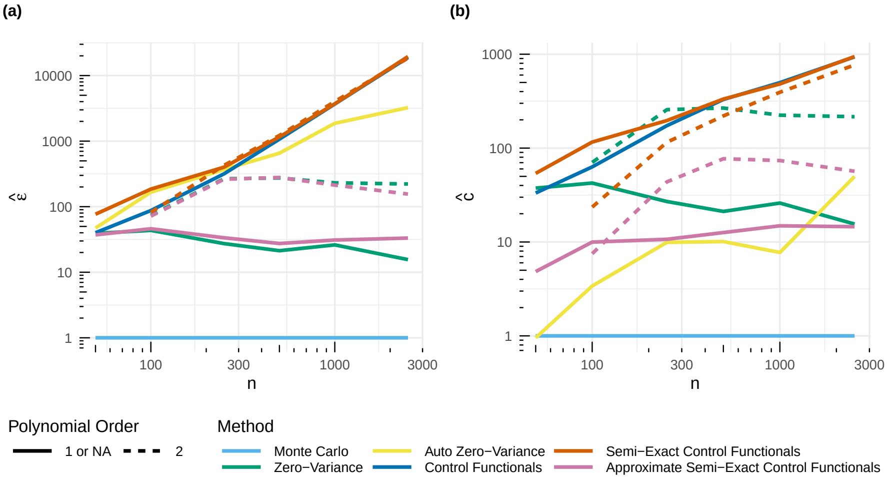

Fig. S3: Recapture example (a) estimated *statistical efficiency* and (b) estimated *computational efficiency* when the unadjusted Langevin algorithm is used in place of the Metropolis-adjusted Langevin algorithm. Efficiency here is reported as an average over the 11 expectations of interest.

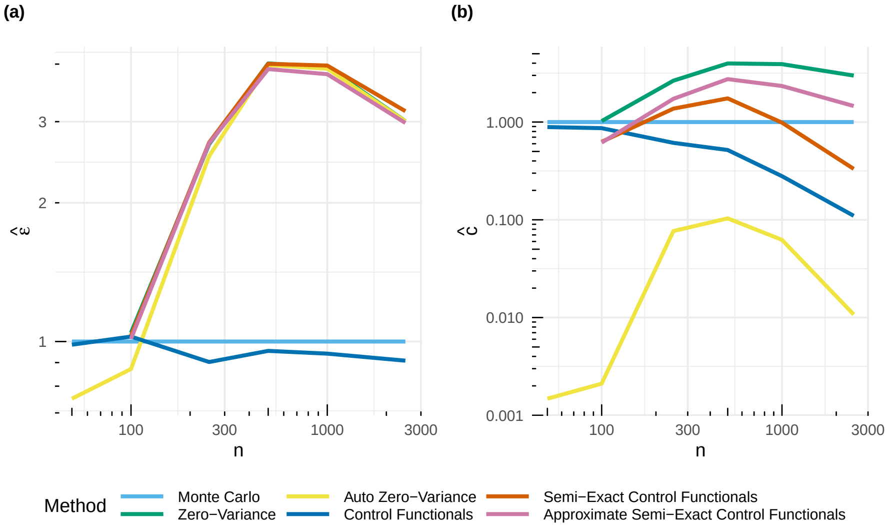

Fig. S4: Sonar example (a) estimated *statistical efficiency* and (b) estimated *computational efficiency* when the unadjusted Langevin algorithm is used in place of the Metropolis-adjusted Langevin algorithm.

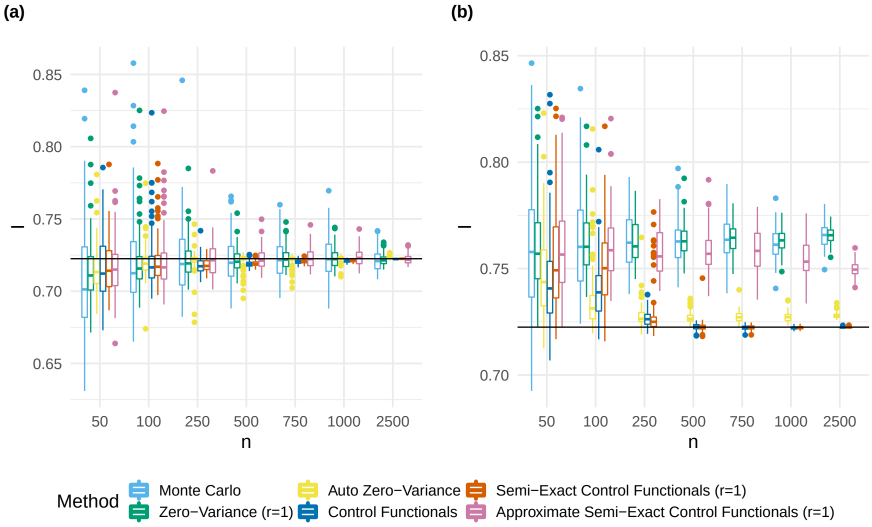

Fig. S5: Recapture example (a) boxplots of 100 estimates of $\int x_1 P_{x|y} \mathrm{d}x$ when the Metropolis-adjusted Langevin algorithm is used for sampling and (b) boxplots of 100 estimates of $\int x_1 P_{x|y} \mathrm{d}x$ when the unadjusted Langevin algorithm is used for sampling. The black horizontal line represents the gold standard of approximation.

10. Reproducing Kernels and Worst-Case Error

The purpose of this section is to review some basic results about worst-case error analysis in a reproducing kernel Hilbert space context. In Appendices 11 and 12 these results are used to prove Proposition 1 and Theorem 1.

Let $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive-definite kernel such that $\int |k(x,y)| p(x)\,\mathrm{d}x < \infty$ for every $y \in \mathbb{R}^d$ and $\mathcal{H}(k)$ the reproducing kernel Hilbert space of k. The *worst-case error* in $\mathcal{H}(k)$ of any weights $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ and any distinct points $\{x^{(i)}\}_{i=1}^n \subset \mathbb{R}^d$ is defined as

$$e_{\mathcal{H}(k)}(v; \{x^{(i)}\}_{i=1}^n) := \sup_{\|h\|_{\mathcal{H}(k)} \leq 1} \left| \int h(x)p(x)\,\mathrm{d}x - \sum_{i=1}^n v_i h(x^{(i)}) \right|. \tag{S15}$$

In this appendix we consider a fixed set of points $\{x^{(i)}\}_{i=1}^n$ and employ the shorthand $e_{\mathcal{H}(k)}(v)$ for $e_{\mathcal{H}(k)}(v; \{x^{(i)}\}_{i=1}^n)$. Then a standard result (see, for example, Section 10.2 in Novak & Woźniakowski, 2010) is that the worst-case error admits a closed form

$$e_{\mathcal{H}(k)}(v) = \left(\int\int k(x,y)p(x)p(y)\,\mathrm{d}x\,\mathrm{d}y - 2\sum_{i=1}^n v_i \int k(x, x^{(i)})p(x)\,\mathrm{d}x + v^\top K v \right)^{1/2}, \tag{S16}$$

where K is the $n \times n$ matrix with entries $[K]_{i,j} = k(x^{(i)}, x^{(j)})$, and

$$\left| \int h(x)p(x)\,\mathrm{d}x - \sum_{i=1}^n v_i h(x^{(i)}) \right| \leq \|h\|_{\mathcal{H}(k)}\, e_{\mathcal{H}(k)}(v) \tag{S17}$$

for any $h \in \mathcal{H}(k)$. Because the worst-case error in (S16) can be written as the quadratic form

$$e_{\mathcal{H}(k)}(v) = (k_{pp} - 2v^\top k_p + v^\top K v)^{1/2},$$

where $k_{pp} = \int\int k(x,y)p(x)p(y)\,\mathrm{d}x\,\mathrm{d}y$ and $[k_p]_i = \int k(x, x^{(i)})p(x)\,\mathrm{d}x$, the weights v which minimise it take an explicit closed form:

$$v_{\text{opt}} = \operatorname*{arg\,min}_{v \in \mathbb{R}^n} e_{\mathcal{H}(k)}(v) = K^{-1} k_p$$

Let $\Psi = \{\psi_0, \ldots, \psi_{m-1}\}$ be a collection of $m \leq n$ basis functions for which the generalised Vandermonde matrix

$$P_\Psi = \begin{bmatrix} \psi_0(x^{(1)}) & \cdots & \psi_{m-1}(x^{(1)}) \\ \vdots & \ddots & \vdots \\ \psi_0(x^{(n)}) & \cdots & \psi_{m-1}(x^{(n)}) \end{bmatrix},$$

has full rank. In this paper we are interested in weights which satisfy the semi-exactness conditions $\sum_{i=1}^n v_i \psi(x^{(i)}) = \int \psi(x)p(x)\,\mathrm{d}x$ for every $\psi \in \Psi$. Minimising the worst-case error under these constraints gives rise to the weights

$$v_{\text{opt}}^\Psi = \operatorname*{arg\,min}_{v \in \mathbb{R}^n} e_{\mathcal{H}(k)}(v) \quad \text{s.t.} \quad \sum_{i=1}^n v_i \psi(x^{(i)}) = \int \psi(x)p(x)\,\mathrm{d}x \quad \text{for every } \psi \in \Psi. \tag{S18}$$

These weights can be solved from the linear system (Karvonen et al., 2018, Theorem 2.7 and Remark D.1)

$$\begin{bmatrix} K & P_\Psi \\ P_\Psi^\top & 0 \end{bmatrix} \begin{bmatrix} v_{\text{opt}}^\Psi \\ a \end{bmatrix} = \begin{bmatrix} k_p \\ \psi_p \end{bmatrix},$$

where $a \in \mathbb{R}^q$ is a nuisance vector and the ith element of ψ_p is $\int \psi_{i-1}(x) p(x) \, \mathrm{d}x$. Note that (S18) is merely a quadratic programming problem under the linear equality constraint $P_\Psi^\top v = \psi_p$.

These facts will be used in Appendices 11 and 12 to prove Proposition 1 and Theorem 1. Their relevance derives from the fact that $e_{\mathcal{H}(k_0)}(v; \{x^{(i)}\}_{i=1}^n)$ coincides with the kernel Stein discrepancy between p and the discrete measure $\sum_{i=1}^n v_i \delta(x^{(i)})$.

11. Proof of Proposition 1

The following proof relies on the results reviewed in Appendix 10.

Proof of Proposition 1. Applying the results reviewed in Appendix 10 with $k = k_0$ and $\psi_j = \mathcal{L}\phi_j$, for which $k_p = 0$ and $\psi_p = e_1$ from (9), we see that the solution to the optimisation problem

$$v_{\text{opt}}^{\mathcal{F}} = \operatorname*{arg\,min}_{v \in \mathbb{R}^n} e_{\mathcal{H}(k_0)}(v; \{x^{(i)}\}_{i=1}^n) \quad \text{s.t.} \quad \sum_{i=1}^n v_i h(x^{(i)}) = \int h(x) p(x) \, \mathrm{d}x \quad \text{for every } h \in \mathcal{F}$$

can be obtained by solving the linear system

$$\begin{bmatrix} K_0 & P \\ P^\top & 0 \end{bmatrix} \begin{bmatrix} v_{\text{opt}}^{\mathcal{F}} \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ e_1 \end{bmatrix}.$$

A straightforward application of the block matrix inversion formula then gives

$$v_{\text{opt}}^{\mathcal{F}} = K_0^{-1} P (P^\top K_0^{-1} P)^{-1} e_1 = w,$$

where in the final equality we have recognised this expression as being identical to the weights w used in our semi-exact control functional method, i.e. $I_{\text{SECF}}(f) = \sum_{i=1}^n w_i f(x^{(i)})$ by (14) and (15). By (9) the only non-zero element on the right-hand side of (S16) is $v^\top K_0 v$. Thus we have characterised the weights $\boldsymbol{w}$ in the semi-exact control functional method as the solution to the problem

$$w = \operatorname*{arg\,min}_{v \in \mathbb{R}^n} (v^\top K_0 v)^{1/2} \quad \text{s.t.} \quad \sum_{i=1}^n v_i h(x^{(i)}) = \int h(x) p(x) \, \mathrm{d}x \quad \text{for every } h \in \mathcal{F}. \tag{S19}$$

If $f = h + g$ with $h \in \mathcal{F}$ and $g \in \mathcal{H}(k_0)$, then it follows from the integral semi-exactness property (S19) that

$$|I(f) - I_{\text{SECF}}(f)| = |I(g) - I_{\text{SECF}}(g) + I(h) - I_{\text{SECF}}(h)| = |I(g) - I_{\text{SECF}}(g)|.$$

Applying (S17) and (S19) yields

$$|I(f) - I_{\text{SECF}}(f) \le \|g\|_{\mathcal{H}(k_0)} \, e_{\mathcal{H}(k_0)}(w; \{x^{(i)}\}_{i=1}^n) = \|g\|_{\mathcal{H}(k_0)} \, (w^\top K_0 w)^{1/2}.$$

Since this bound is valid for any decomposition $f = h + g$ with $h \in \mathcal{F}$ and $g \in \mathcal{H}(k_0)$ we have

$$|I(f) - I_{\mathrm{SECF}}(f)| \;\leq\; \inf_{\substack{f=h+g \\ h\in\mathcal{F},\, g\in\mathcal{H}(k_0)}} \|g\|_{\mathcal{H}(k_0)}(w^\top K_0 w)^{1/2} \;=\; |f|_{k_0,\mathcal{F}}\,(w^\top K_0 w)^{1/2}$$

as claimed. □

12. Proof of Theorem 1

The following proof relies on the worst-case error results reviewed in Appendix 10, together with the following result, due to Hodgkinson et al. (2020), which studies the convergence of the worst-case error (i.e. the kernel Stein discrepancy) of a weighted combination of the states $\{x^{(i)}\}_{i=1}^n$, where the weights $\tilde{w}$ are obtained by minimising the worst-case error subject to a non-negativity constraint:

Theorem S1. *Let p be a probability density on $\mathbb{R}^d$ and $k_0 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a reproducing kernel which satisfies*

$$\int k_0(x,y)p(x)\,\mathrm{d}x = 0$$

for every $y \in \mathbb{R}^d$. Let q be a probability density with $p/q > 0$ on $\mathbb{R}^d$ and consider a q-invariant Markov chain $(\boldsymbol{x}^{(i)})_{i=1}^n$, assumed to be V-uniformly ergodic for some $V : \mathbb{R}^d \to [1,\infty)$, such that

$$\sup_{x\in\mathbb{R}^d} V(x)^{-r} \left\{\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right\}^4 k_0(x,x)^2 < \infty$$

for some $0 < r < 1$. Let

$$\tilde{w} = \operatorname*{arg\,min}_{v\in\mathbb{R}^n} e_{\mathcal{H}(k_0)}(v; \{x^{(i)}\}_{i=1}^n) \quad \textit{s.t.} \quad \sum_{i=1}^n v_i = 1 \quad \textit{and} \quad v \geq 0. \tag{S20}$$

Then $e_{\mathcal{H}(k_0)}(\tilde{w}; \{x^{(i)}\}_{i=1}^n) = O_P(n^{-1/2})$.

Proof. A special case of Theorem 1 in Hodgkinson et al. (2020). □

The sense in which Theorem S1 will be used is captured in the following corollary, which follows from the observation that removal of the non-negativity constraint in (S20) does not increase the worst-case error:

Corollary S1. *Under the same hypotheses as Theorem S1, let*

$$\bar{w} = \operatorname*{arg\,min}_{v\in\mathbb{R}^n} e_{\mathcal{H}(k_0)}(v; \{x^{(i)}\}_{i=1}^n) \quad \textit{s.t.} \quad \sum_{i=1}^n v_i = 1. \tag{S21}$$

Then $e_{\mathcal{H}(k_0)}(\bar{w}; \{x^{(i)}\}_{i=1}^n) \leq e_{\mathcal{H}(k_0)}(\tilde{w}; \{x^{(i)}\}_{i=1}^n) = O_P(n^{-1/2})$.

Now the proof of Theorem 1 can be presented:

Proof of Theorem 1. From Assumption A2 and Lemma 4 we have $(\boldsymbol{P}^\top \boldsymbol{K}_0^{-1} \boldsymbol{P})^{-1}$ is almost surely well-defined. In the proof of Proposition 1 we saw that

$$e_{\mathcal{H}(k_0)}(w; \{x^{(i)}\}_{i=1}^n)^2 = w^\top K_0 w$$

and

$$(I(f) - I_{\mathrm{SECF}}(f))^2 \leq |f|_{k_0,\mathcal{F}}^2\, w^\top K_0 w$$

Plugging in the expression for $w = K_0^{-1}P(P^\top K_0^{-1}P)^{-1}e_1$ in (15), we obtain

$$e_{\mathcal{H}(k_0)}(w; \{x^{(i)}\}_{i=1}^n)^2 = [(P^\top K_0^{-1}P)^{-1}]_{1,1} \tag{S22}$$

and

$$(I(f) - I_{\mathrm{SECF}}(f))^2 \leq |f|_{k_0,\mathcal{F}}^2 [(P^\top K_0^{-1}P)^{-1}]_{1,1}.$$

It therefore suffices to consider the stochastic fluctuations of $[(P^\top K_0^{-1}P)^{-1}]_{11}$ as $n \to \infty$. To this end, let $[\boldsymbol{\Psi}]_{i,j} := \mathcal{L}\phi_j(\boldsymbol{x}^{(i)})$ and consider the block matrix

$$\frac{\boldsymbol{P}^\top \boldsymbol{K}_0^{-1} \boldsymbol{P}}{\mathbf{1}^\top \boldsymbol{K}_0^{-1}\mathbf{1}} = \begin{bmatrix} 1 & \frac{\mathbf{1}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi}}{\mathbf{1}^\top \boldsymbol{K}_0^{-1}\mathbf{1}} \\ \frac{\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{K}_0^{-1}\mathbf{1}} & \frac{\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi}}{\mathbf{1}^\top \boldsymbol{K}_0^{-1}\mathbf{1}}. \end{bmatrix} \tag{S23}$$

From the block matrix inversion formula we have

$$\begin{aligned} \left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{K}_0^{-1}\boldsymbol{P}}{\mathbf{1}^\top \boldsymbol{K}_0^{-1}\mathbf{1}}\right)^{-1}\right]_{1,1} &= \left[1 - \frac{\mathbf{1}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{K}_0^{-1}\mathbf{1}}\right]^{-1} \\ &= \left[1 - \frac{\langle \mathbf{1}, \Pi\mathbf{1}\rangle_n}{\langle \mathbf{1}, \mathbf{1}\rangle_n}\right]^{-1}. \end{aligned} \tag{S24}$$

Since $\Pi = \boldsymbol{\Psi}(\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}$ and

$$\begin{aligned} \|\Pi\mathbf{1}\|_n^2 &= \mathbf{1}^\top \Pi^\top \boldsymbol{K}_0^{-1}\Pi\mathbf{1} \\ &= \mathbf{1}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\mathbf{1} \\ &= \mathbf{1}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top \boldsymbol{K}_0^{-1}\mathbf{1} \\ &= \mathbf{1}^\top \boldsymbol{K}_0^{-1}\Pi\mathbf{1} \\ &= \langle \mathbf{1}, \Pi\mathbf{1}\rangle_n, \end{aligned}$$

our Assumption A3 implies that (S24) is almost surely asymptotically bounded, say by a constant $C \in [0, \infty)$. In other words, it almost surely holds that

$$[(P^\top K_0^{-1}P)^{-1}]_{1,1} \;\leq\; C(\mathbf{1}^\top \boldsymbol{K}_0^{-1}\mathbf{1})^{-1}$$

for all sufficiently large n.

To complete the proof we evoke Corollary S1, noting that the weights $\bar{w}$ defined in Corollary S1 satisfy $e_{\mathcal{H}(k_0)}(\bar{w}; \{x^{(i)}\}_{i=1}^n) = (1^\top K_0^{-1}1)^{-1/2}$, which follows from (S22) with $P = 1$. Thus from Corollary S1 we conclude

$$[(1^\top K_0^{-1}1)^{-1}]_{1,1}^{1/2} = e_{\mathcal{H}(k_0)}(\bar{w}; \{x^{(i)}\}_{i=1}^n) \leq e_{\mathcal{H}(k_0)}(\tilde{w}; \{x^{(i)}\}_{i=1}^n) = O_P(n^{-1/2}),$$

as required.

References

Adams, R. A. & Fournier, J. J. (2003). *Sobolev Spaces.* Elsevier.

Alaoui, A. & Mahoney, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama & R. Garnett, eds., vol. 28. Curran Associates, Inc.

Berlinet, A. & Thomas-Agnan, C. (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Springer Science & Business Media.

CHWIALKOWSKI, K., STRATHMANN, H. & GRETTON, A. (2016). A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan & K. Q. Weinberger, eds., vol. 48 of *Proceedings of Machine Learning Research*. New York, New York, USA: PMLR.

ERMAK, D. L. (1975). A computer simulation of charged particles in solution. I. Technique and equilibrium properties. *The Journal of Chemical Physics* **62**, 4189–4196.

FASSHAUER, G. E. & YE, Q. (2011). Reproducing kernels of generalized Sobolev spaces via a Green function approach with distributional operators. *Numerische Mathematik* **119**, 585–611.

GARREAU, D., JITKRITTUM, W. & KANAGAWA, M. (2017). Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269* .

HODGKINSON, L., SALOMONE, R. & ROOSTA, F. (2020). The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv preprint arXiv:2001.09266* .

KARVONEN, T., OATES, C. J. & SÄRKKÄ, S. (2018). A Bayes–Sard cubature method. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, vol. 31.

LIU, Q., LEE, J. & JORDAN, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan & K. Q. Weinberger, eds., vol. 48 of *Proceedings of Machine Learning Research*. New York, New York, USA: PMLR.

NOVAK, E. & WOŹNIAKOWSKI, H. (2010). *Tractability of Multivariate Problems. Volume II: Standard Information for Functionals*, vol. 12 of *EMS Tracts in Mathematics*. European Mathematical Society.

OATES, C. J., GIROLAMI, M. & CHOPIN, N. (2017). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 695–718.

PARISI, G. (1981). Correlation functions and computer simulations. *Nuclear Physics B* **180**, 378–384.

RUDI, A., CAMORIANO, R. & ROSASCO, L. (2015). Less is more: Nyström computational regularization. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press.

RUDI, A., CARRATINO, L. & ROSASCO, L. (2017). FALKON: An optimal large scale kernel method. In *Proceedings of the 31st Conference on Neural Information Processing Systems*. Curran Associates Inc.

SMOLA, A. J. & SCHÖKOPF, B. (2000). Sparse greedy matrix approximation for machine learning. In *Proceedings of the 17th International Conference on Machine Learning*, ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

STEINWART, I. & CHRISTMANN, A. (2008). *Support Vector Machines*. Information Science and Statistics. Springer.

WILLIAMS, C. K. I. & SEEGER, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich & V. Tresp, eds., vol. 13. MIT Press.

[*Received on 2 January 2017. Editorial decision on 1 April 2017*]