

---

# Kernel von Mises Formula of the Influence Function

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 The influence function (also known as the first variation and Fisher-Rao gradient)  
2 of a statistical functional is the Riesz representer of its derivative operator. It is  
3 the key analytic object in both the theory and implementation of estimators in  
4 semiparametric statistical and machine learning models, e.g., one-step estimator,  
5 targeted learning, debiased machine learning. It is also essential for inference  
6 about their statistical, robustness and interpretability properties, e.g., for finding  
7 confidence intervals, partial identification and misspecification bounds, adversarial  
8 perturbations, influential data points. However, the analytic derivation of the influ-  
9 ence function is often an obstruction to the broader adoption of these methods by  
10 practitioners. Toward automating this task, we derive a regularized representation  
11 of the influence function using spectral theory of positive semidefinite kernels.  
12 Based on this representation we construct an estimator that: (i) is a nonparametric  
13 functional RKHS estimator; (ii) admits theoretical guarantees in function norms  
14 relevant for downstream tasks; (iii) can be computed via automatic differentiation  
15 or finite differences, without requiring analytic derivation by the user.

## 16 1 Introduction

17 The target of an estimation procedure or learning algorithm often takes the form of a functional  
18 of the probability distribution governing the observed data and latent variables of a statistical or  
19 machine learning (ML) model. The influence function of such a functional is an analytic object that  
20 characterizes the theoretical properties of its estimators in relationship to the model. Beyond its  
21 foundational role in the theory, the influence function is a key requirement in the construction of  
22 efficient and debiased estimators, including the one-step adaptive estimator [Bic82], targeted learning  
23 [VR06; Cho+24], debiased machine learning [Che+22]. It is also essential for constructing statistical  
24 inference procedures for these estimators [Mis47; KL76; Cha86; Kla87; Vaa91; New94]. Moreover, the influence  
25 function has been used to study: (i) partial identification and robustness to misspecification [HM95;  
26 Muk19; Sem20; Muk21; CC23; Sem25]; (ii) influence of individual observations and highly influential  
27 observations [Hub92; Mad+17; Pru+20; BGM20]; (iii) interpretability of econometric and machine learning  
28 models [AGS17; Muk18; AGS20a; AGS20b; KL17; Gro+23]. Despite its broad utility, the adoption of influence-  
29 function-based methods often hinges on the user’s capacity to derive the influence function analytically  
30 for their specific application. This task can be time-consuming and requires familiarity with functional  
31 analysis and often highly specialized technical knowledge, posing a significant obstruction to broad  
32 adoption [CLV19; Hin+22; Ken24; JWZ22].

### 33 1.1 Influence function and von Mises formula

34 Let  $\mathcal{X} \subset \mathbb{R}^d$  be a sample space and  $\mathcal{P}$  denote a collection of probability measures on  $\mathcal{X}$ . We  
35 will assume that  $\mathcal{P}$  is a nonparametric model with  $L_0^2(P)$  tangent spaces [Bic+93], [Vaa00, ch25]. Let  
36  $\theta : \mathcal{P} \rightarrow \mathbb{R}$  denote a known statistical functional and let  $X_1, \dots, X_n$  be an i.i.d. sample from the  
37 unknown data generating distribution  $P \in \mathcal{P}$ . We consider parameters  $\theta(P)$  that can be estimated at

the parametric  $\sqrt{n}$  rate and admit estimators  $\hat{\theta}(X_1, \dots, X_n)$  that are asymptotically linear:

$$\sqrt{n}(\hat{\theta} - \theta(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_P(X_i) + o_P(1). \quad (1.1)$$

In this representation,  $\psi_P : \mathcal{X} \rightarrow \mathbb{R}$  is an  $L_0^2(P)$  function, so that  $E \psi(X)^2 < +\infty$  and  $E \psi(X) = 0$ , and the  $o_P(1)$  term vanishes as  $n \rightarrow \infty$  in probability. The map  $\psi_P(x)$  is known as the influence function of the estimator  $\hat{\theta}$ . According to (1.1), the contribution of an observation  $X_i$  to the fluctuations in the estimator is approximately  $\psi(X_i)/n$ , and the variance of the estimator can be estimated by  $E \psi(X)^2/n$ . Furthermore, a preliminary estimate  $\theta'$  can be improved by an update  $\theta''$  that sets  $E \psi_P(X) \approx 0$ . Implementing these and related methods requires the function  $x \mapsto \psi_P$  as input.

Asymptotic linearity (1.1) is closely related (roughly equivalent) to the differentiability of the functional  $\theta(P)$  with respect to perturbations of the measure  $P$  [Kla87; Vaa91; FS19]. In regular cases, the influence function of an estimator  $\hat{\theta}(X_1, \dots, X_n)$  coincides with the  $L^2(P)$  gradient of the derivative of the functional  $\theta(P)$  [New94; CS18]. The following calculation extends those of [Mis47; IN22] and allows to obtain evaluations of the influence function  $\psi_P$  at a point  $x \in \mathcal{X}$  from evaluations of the (known) functional  $\theta$  on certain perturbations of the measure  $P$ . A useful analogy is to think of computing the partial derivatives for a multivariate function of  $\mathbb{R}^2$  or  $\mathbb{R}^3$  to evaluate the gradient vector.

**THEOREM 1.1** [von Mises formula]. *Let  $\theta : \mathcal{P} \rightarrow \mathbb{R}$  be a pathwise differentiable functional on the nonparametric model  $\mathcal{P}$  with influence functions  $\psi_P \in L_0^2(P)$  for  $P \in \mathcal{P}$ . Suppose  $P$  is an absolutely continuous probability measure with respect to the Lebesgue measure on  $\mathbb{R}^d$  with a continuous density function  $f$ . Let  $K$  be a bounded probability density function with support in the unit ball  $\{|x| \leq 1\} \subset \mathbb{R}^d$ , and define the dilated kernels by  $K^\delta(x) := \delta^{-d} K(\delta^{-1}x)$  for  $\delta > 0$  and, finally, translate to the location of approximation  $z \in \mathbb{R}^d$  and control the likelihood ratio with  $f(x)$  via a cutoff:*

$$K^{\delta,z}(x) := c K^\delta(z - x) \cdot \mathbf{1}_{\{f > \delta\}}(x), \quad c = \left[ \int_{\{f > \delta\}} K^\delta(z - x) \, dx \right]^{-1}.$$

For  $\delta > 0$  small and  $z \in \{f > 0\} \subset \mathcal{X}$ , consider the family, indexed by the regularization parameter  $\delta$ , of paths  $\{P_t^{\delta,z}\}_{-\epsilon < t \leq 1}$  with parameter  $t$  and density

$$f_t^{\delta,z}(x) := (1 - t)f(x) + tK^{\delta,z}(x), \quad x \in \mathcal{X}.$$

Note that these paths perturb the measure  $P = P_{t=0}$  toward the point-mass distribution at  $z \in \mathcal{X}$ , regularized via the approximation to the identity  $K^\delta$ . Then the following influence function formula holds:

$$\psi_P(z) = \lim_{\delta \rightarrow 0} \left[ \frac{d}{dt} \theta(P_t^{\delta,z}) \right]_{|t=0} \quad (1.2)$$

for  $P$ -almost every  $z \in \mathbb{R}^d$ .

*Proof.* We outline the main ideas of the proof as a way of introducing pathwise differentiability and influence functions and provide the details the Appendix. The score function (derivative of log-density) of the path  $t \mapsto P_t^{\delta,z}$  at  $P$  is

$$\phi_{\delta,z}(x) = \frac{d}{dt} \Big|_{t=0} \log \left\{ f(x) + t[K^\delta(z - x) - f(x)] \right\} = K^\delta(z - x)/f(x) - 1.$$

The score  $\phi_{\delta,z}(x)$  is an  $L_0^2(P)$  function that measures the infinitesimal change in the density at  $x$  as  $P$  is perturbed along the path  $P_t^{\delta,z}$ . By pathwise differentiability of  $\theta$  at  $P$ , the derivative of the functional  $\theta$  along the curve  $t \mapsto P_t^{\delta,z}$  exists and is a bounded linear functional of the score  $\phi_{\delta,z}$ . By the Riesz representation theorem [SS09, 4.5], [Dud18, 5.5.1] for bounded functionals on the Hilbert space  $L_0^2(P)$ , the derivative  $D\theta_P[\phi_{\delta,z}]$  is given by the  $L_0^2(P)$  inner product of the score  $\phi_{\delta,z}$  with the influence function  $\psi_P$ :

$$\frac{d}{dt} \Big|_{t=0} \theta(P_t^{\delta,z}) = D\theta_P[\phi_{\delta,z}] = \int_{\text{spt } P} \psi_P(x) K^\delta(z - x) \, dx = (\psi_P * K^\delta)(z).$$

74 The assumed properties of the mollification kernels  $K^\delta$  ensure that it is an approximation to the  
75 identity [SS09, 3.2] in the sense that it converges as  $\delta \rightarrow 0$  to the singular point mass distribution in  
76 integral pairing with a  $L^1_{\text{loc}}(\mathbb{R}^d)$  function. By the Lebesgue differentiation theorem [SS09, 3.3], [Dud18,  
77 7.2] it follows that the convolution

$$(\psi_P * K^\delta)(z) \rightarrow \psi_P(z) \quad \text{as } \delta \rightarrow 0$$

78 converges pointwise at the Lebesgue points of  $\psi_P$  and therefore for  $P$ -almost every  $z \in \mathbb{R}^d$ .  $\square$

79 Let's interpret Theorem 1.1. It says that to compute a single value of the influence function, it  
80 is sufficient to compute the values of the functional  $\theta$  along a certain perturbation to  $P$ . This is  
81 natural, given that the influence function encodes the sensitivity of  $\theta$  to *all admissible perturbations*  
82 of  $P$ . Provided that we have a device for computing the derivative  $d\theta/dt$  in (1.2) numerically, this  
83 representation can be used to numerically query the unknown function  $\psi_P$ . [CLV19; JWZ22] use finite  
84 differences with a similar von Mises representation (1.2) for numerical approximation of  $\psi_P(x)$ . Our  
85 regularity assumptions for this result are different from those in the literature, by employing Lebesgue  
86 differentiation we make no additional regularity assumption about  $\psi_P$ .

87 However, in statistical applications, one typically requires the entire map  $z \mapsto \psi(z)$  rather than a  
88 particular value  $\psi(z)$ . For example, to find the influential data points for estimating  $\theta$ , one seeks  
89 the global maximum or level sets of  $\psi$ ; for constructing a debiased estimator of  $\theta$  one needs to  
90 integrate against  $\psi$ ; to find the influential data points in the Wasserstein sense, one needs to compute a  
91 differential operator of the gradient  $\nabla_x \psi(x)$  and maximize the result. Therefore, in practice, formula  
92 (1.2) is used to evaluate *many* values of  $\psi$  *simultaneously*. With this in mind, we note that (1.2)  
93 requires a separate computation for each evaluation and that the perturbations toward a point-mass  
94 have been found to be numerically challenging [CLV19; JWZ22]; we also note that the regularization  
95 in Theorem 1.1 does not take into account properties of the measure  $P$  such as concentration or  
96 properties of the function  $\psi$  such as smoothness. These observations suggest that (1.2) may not be  
97 statistically and numerically optimal for estimating the entire function  $\psi$  or even isolated values of  $\psi$ .

98 **Contribution** With the goal of formulating a functional estimator of the influence function  $\psi$ ,  
99 we derive a spectral representation of  $\psi$  in terms of pathwise derivatives of  $\theta$  along well-behaved  
100 perturbations of  $P$ . Specifically, one that provides approximation to all values of  $\psi$  simultaneously,  
101 while requiring only a few pathwise derivatives of  $\theta$  for a low-rank approximation. We accomplish  
102 this by constructing nonparametric principle components (PCA) of the model  $\mathcal{P}$  locally at  $P$  and use  
103 them as a complete and ordered basis of perturbations to  $P$  that span the tangent space  $L^2_0(P)$  and,  
104 in particular,  $\psi_P$ . With this basis, we derive a regularized spectral von Mises representation of  $\psi_P$ ,  
105 which leads to a low-rank functional approximation in terms of a small number of pathwise derivatives  
106 of  $\theta$  along the leading principle components of  $\mathcal{P}$  at  $P$ . We then use our spectral representation with  
107 the eigenfunctions of a universal Mercer kernel. We use the Nyström methods for integral operators  
108 to estimate the kPCA and prove consistency of the resulting estimator of  $\psi_P$  by a low-rank projection  
109 on the balls of a reproducing kernel Hilbert space (rkHs). We will explore the rates of convergence  
110 and computational aspects of our estimator in future work.

## 111 2 Spectral von Mises representation

### 112 2.1 Exact representation

113 We begin by finding a variational representation of the influence function in terms of pathwise  
114 derivatives of the parameter  $\theta$ . The following lemma is an immediate consequence of the Riesz  
115 representation theorem [Dud18, 5.5.1] and Cauchy-Schwarz inequality [Dud18, 5.1.4] and records in  
116 a suitable form the basic observation: the influence function  $\psi_P$  is the direction of most rapid  
117 perturbation to the measure  $P$  for the value of the functional  $\theta(P)$  in the Fisher-Rao geometry (with  
118 the  $L^2(P)$  metric tensor) of the model  $\mathcal{P}$ .

119 **LEMMA 2.1.** *Let  $\theta$  be a pathwise differentiable functional on  $\mathcal{P}$  with derivative operator  $D\theta_P$  and*  
120 *influence function  $\psi_P$  for  $P \in \mathcal{P}$ . Then the influence function is the unique solution to the following*

121 *functional optimization programs:*

$$\begin{aligned}\psi_P &= -\arg \min_{\phi \in L_0^2(P)} \left\{ D\theta_P[\phi] ; \|\phi\|_{L^2(P)} \leq 1 \right\} \\ &\propto -\arg \min_{\phi \in L_0^2(P)} \left\{ D\theta_P[\phi] + \lambda_p \|\phi\|_{L^2(P)}^2 \right\}, \quad \lambda_p > 0.\end{aligned}\tag{2.1}$$

122 *The proportionality constant in (2.1) is 1 if and only if the penalty loading is  $\lambda_p = 1/2$ .*

## 123 2.2 Regularized representation

124 In (2.1) we used the duality between constrained and penalized optimization. For the exact representation of  $\psi$ , both the constraint and the penalty are in terms of the  $L^2(P)$  norm (i.e., the metric tensor of the Fisher-Rao distance on  $\mathcal{P}$ ), which gives rise to the geometry where the influence function is the gradient perturbation. This exact variational representation suggests a strategy for constructing a regularized approximation of  $\psi$  as follows. Suppose there is a function space  $H \subset L_0^2(P)$  with a norm  $\|\phi\|_H$  that quantifies a suitable notion of smoothness of functions  $\phi \in H$ . Suppose we wish to find the best approximation of  $\psi$  in  $H$  with a given degree of smoothness as measures by  $\|\cdot\|_H$ . For example,  $H$  can be a Sobolev space or a space of splines. Then the projection  $\psi_M$  of  $\psi$  on the ball

$$B_M := \{\phi ; \|\phi\|_H \leq M\} \subset H \subset L_0^2(P)$$

132 of radius  $M > 0$  is the desired approximation, and  $M$  controls the degree of regularization. If  $H$  is dense in  $L_0^2(P)$ , then we indeed obtain an approximation of  $\psi$  by the projection  $\psi_M$  that improves and converges to  $\psi$  as  $M \rightarrow \infty$ .

135 **LEMMA 2.2.** *Let  $\theta$  be a pathwise differentiable functional on  $\mathcal{P}$  with derivative operator  $D\theta_P$  and influence function  $\psi_P$  for  $P \in \mathcal{P}$ . Let  $(H, \|\cdot\|_H)$  be a Hilbert space, densely contained in  $L_0^2(P)$ . Then the projection of the influence function  $\psi_P$  on the set  $B_M$  is the unique solution to the following functional optimization programs:*

$$\begin{aligned}\psi_{P,M} &:= -\arg \min_{\phi \in H} \left\{ D\theta_P[\phi] ; \|\phi\|_{L^2(P)} \leq 1 \text{ and } \|\phi\|_H \leq M \right\} \\ &= -\arg \min_{\phi \in H} \left\{ D\theta_P[\phi] + 1/2 \|\phi\|_{L^2(P)}^2 + \lambda_r \|\phi\|_H^2 \right\} \quad =: \psi_{P,\lambda}\end{aligned}\tag{2.2}$$

139 *for some regularization loading  $\lambda_r = \lambda_r(M) \geq 0$ . Furthermore,  $\psi_M \rightarrow \psi$  in  $L_0^2(P)$  as  $M \rightarrow \infty$  and, equivalently,  $\psi_\lambda \rightarrow \psi$  in  $L_0^2(P)$  as  $\lambda_r \rightarrow 0$ .*

## 141 2.3 Spectral representation

142 We obtained a regularized functional representation (2.2) of the influence function in terms of the evaluation of the pathwise derivative of the parameter  $\theta$ :

$$D\theta_P[\phi] = \frac{d}{dt} \Big|_{t=0} \theta(P_t^\phi), \quad \text{where} \quad \frac{d}{dt} \Big|_{t=0} \log P_t^\phi = \phi, \quad P_{t=0} = P,$$

144 and the path  $\{P_t^\phi\}_{0 \leq t < \epsilon}$  can be taken to be any regular parametric model with parameter  $t$  and score function  $\phi \in H$  at  $P$ . This representation of  $\psi_P$  is rather implicit. But it is the correct representation because it emphasizes the geometry of the problem and lends to thinking about  $\psi_P$  as a vector in an inner product space rather than a bag of numbers, one for each  $x \in \mathcal{X}$ . Indeed, going back to our analogy with computing the gradient of a multivariate function on  $\mathbb{R}^3$ , we make the main basic observation of this paper:

150 **Main idea:** *The mathematically fruitful way of thinking about partial derivatives of  $\theta$  in  $\mathcal{P}$  is not along perturbations toward a point mass at each  $x \in \mathcal{X}$ , but rather along the directions of an orthonormal basis on the tangent space  $L_0^2(P)$ .*

153 Equipped with this intuition, we solve the optimization problem (2.2) analytically to obtain an infinite Fourier series representation of the regularized influence function  $\psi_{P,\lambda}$ . Our main result is:

155 THEOREM 2.3 [Spectral von Mises formula]. Suppose there exists an orthonormal basis of functions  
 156  $\{e_j : \mathcal{X} \rightarrow \mathbb{R}\}_{j \geq 1}$  for  $L_0^2(P)$  and a decreasing to zero sequence of scalars  $\{\sigma_j > 0\}_{j \geq 1}$  such that  
 157  $\{\sqrt{\sigma_j} e_j\}_{j \geq 1}$  is an orthonormal basis for the Hilbert space  $H \subset L_0^2(P)$ .

158 Let  $S : L_0^2(P) \rightarrow H$  be the linear regularization operator given in diagonalized form by

$$S[u](x) = \sum_{j=1}^{\infty} \sigma_j \langle u, e_j \rangle_{L^2(P)} e_j(x), \quad u \in L_0^2(P) \quad (2.3)$$

159 and assume that its adjoint operator  $S^* : H \rightarrow L_0^2(P)$  is the inclusion  $S^*[v] = v$ , so that the inner  
 160 products of  $H$  and  $L_0^2(P)$  are related by

$$\langle Su, v \rangle_H = \langle u, S^*v \rangle_{L^2(P)} = \langle u, v \rangle_{L^2(P)}, \quad \text{for all } u \in L_0^2(P), v \in H. \quad (2.4)$$

161 Let  $\theta$  be a pathwise differentiable functional on  $\mathcal{P}$  with derivative operator  $D\theta_P$  and influence  
 162 function  $\psi_P \in L_0^2(P)$  for  $P \in \mathcal{P}$ . Then the following representation holds in the norm of  $H$ :

$$\psi_{P,\lambda}(x) = \lim_{r \rightarrow \infty} \sum_{j=1}^r \frac{1}{1 + 2\lambda/\sigma_j} \left[ \frac{d}{dt} \theta(P_t^j) \right]_{|t=0} e_j(x), \quad (2.5)$$

163 where, for each basis function  $e_j$ , the path  $t \mapsto P_t^j$  can be any regular perturbation of  $P$  with the  
 164 score function  $\partial_t \log P_t^j = e_j$ .

165 Let's interpret Theorem 2.3 and compare with Theorem 1.1. Formula (2.5) is similar to formula (1.2)  
 166 in that it expresses the (regularized) influence function  $\psi_P$  in terms of pathwise derivatives of the  
 167 functional  $\theta(P)$  along certain regular perturbations  $P_t$  of the measure  $P$ . So, in order to compute  
 168 with formula (2.5), one requires the same numerical tools as for implementing formula (1.2). The  
 169 main difference is in the choice of the perturbation directions  $\partial_t \log P_t$  that are employed by the two  
 170 representations: (i) In (1.2), the perturbation depends on the evaluation point  $z \in \mathcal{X}$ . By contrast,  
 171 in (2.5) the scores  $e_j$  are fixed, once the approximating space  $H$  and the data distribution  $P$  are  
 172 fixed; and, once the derivatives  $\{D\theta[e_j]\}_{j=1}^r$  are computed, an approximation to all values of  $\psi_P$  are  
 173 obtained. (ii) The directions of perturbation  $e_j$  depend on the approximating function class  $H$ , which  
 174 allows to adapt to the smoothness of  $\psi$ . (iii) In our proposed implementation,  $H$  is taken to be a  
 175 reproducing kernel Hilbert space (rkHs) of a positive semidefinite kernel  $K$ , and the scores  $e_j$  can be  
 176 interpreted as nonlinear principle components of the measure  $P$ , which allows to adapt to its effective  
 177 dimension. (iv) The principle perturbation directions  $e_j$  of  $P$  are ordered by the magnitude of the  
 178 corresponding multiplier sequence  $\sigma_j$ , which leads to a natural low-rank approximation for  $\psi_\lambda$  by the  
 179 first  $r$  terms of the formula (2.5). (v) The directions  $\{e_j\}_{j=1}^r$  for perturbing  $P$  are well-behaved.

180 *Proof.* We outline the main ideas of the proof and provide the details in the Appendix. Let  $J(\phi)$   
 181 denote the objective function in (2.2) and note that it is strictly convex, so that the first order conditions  
 182 are necessary and sufficient for characterizing the function  $\psi_\lambda \in H$  at which the unique minimum of  
 183  $J$  is attained. For a direction  $v \in H$ , the Gateau derivative of the objective function evaluated at the  
 184 candidate function  $\phi \in H$  is

$$\partial_v J(\phi) = \langle v, \psi \rangle_{L^2(P)} + \langle v, \phi \rangle_{L^2(P)} + 2\lambda \langle v, \phi \rangle_H.$$

185 Applying the adjoint relationship (2.4) to express  $L^2(P)$  inner products in terms of  $H$  inner products,  
 186 obtain

$$\partial_v J(\phi) = \langle v, S\psi \rangle_H + \langle v, S\phi \rangle_H + 2\lambda \langle v, \phi \rangle_H, \quad \phi, v \in H.$$

187 It follows that the  $H$  gradient of  $J$  (the Riesz representer in the  $H$  inner product) is given by

$$\delta_H J(\phi) = S\psi + S\phi + 2\lambda\phi$$

188 and that the solution  $\psi_\lambda$  of (2.2) is characterized by the first order condition  $\delta_H J(\psi_\lambda) = 0$ . Using  
 189 the spectral resolution (2.3) of the smoothing operator  $S$ , the first order condition is equivalent to the  
 190 system of equations

$$\sigma_j \langle \psi, e_j \rangle_{L^2(P)} + \sigma_j \langle \psi_\lambda, e_j \rangle_{L^2(P)} + 2\lambda \langle \psi_\lambda, e_j \rangle_{L^2(P)} = 0, \quad j \geq 1.$$

191 Solving for the Fourier coefficients  $\langle \psi_\lambda, e_j \rangle_{L^2(P)}$  of the spectral resolution of  $\psi_\lambda$  in  $L^2(P)$ , and  
 192 applying Riesz' representation  $D\theta[e_j] = \langle \psi, e_j \rangle_{L^2(P)}$ , obtain formula (2.5).  $\square$

### 3 Regularization in the rkHs of a Mercer kernel

#### 3.1 RKHS setting

A linear space  $H$  of functions  $\phi : \mathcal{X} \rightarrow \mathbb{R}$  with an inner product  $\phi, \varphi \mapsto \langle \phi, \varphi \rangle_H$  is an rkHs if (i) it is complete in the norm  $\|\phi\|_H^2 = \langle \phi, \phi \rangle_H$  of the inner product, and (ii) for every  $x \in \mathcal{X}$ , the evaluation functional  $\phi \mapsto \phi(x)$  is continuous in the topology of the norm of  $H$ . Convergence of a sequence  $\phi_n \rightarrow \phi$  in the norm of  $H$  implies pointwise, and often uniform or stronger, convergence.

According to Riesz' representation theorem [Dud18, t5.5.1], [SS09, t5.3], for every  $x \in \mathcal{X}$ , there exists a function  $k_x \in H$  such that the evaluation functional has the representation  $\phi(x) = \langle \phi, k_x \rangle_H$  for all  $\phi \in H$ . The function  $K : \mathcal{X}^2 \rightarrow \mathbb{R}$  given by  $K(x, y) := k_y(x) = \langle k_y, k_x \rangle_H$  is known as the reproducing kernel of  $H$  and is guaranteed to be a symmetric and positive semidefinite map: the matrix  $[K(x_i, x_j)]_{i,j=1}^n$  is a strictly positive definite for all  $n \geq 1$  and distinct  $\{x_j\} \subset \mathcal{X}$ ; we call any such function a PSD kernel. We call  $k_x(\cdot) = K(\cdot, x)$  the slice of  $K$  at  $x$ . By linearity, any finite superposition of slices  $x \mapsto \sum_{j=1}^n \alpha_j k_{x_j}(x)$ , with  $\{x_j\} \subset \mathcal{X}$  and  $\{\alpha_j\} \subset \mathbb{R}$  and  $n \in \mathbb{N}$ , is in  $H$ . Conversely, according to Moore's theorem [PR16, t2.14], [CS08, ch4], any function  $\phi \in H$  is a (possibly infinite) superposition of kernel slices; and, any PSD kernel  $K$  generates an rkHs of superpositions for which it is the reproducing kernel. In practice, one picks a PSD kernel function and works with the associated rkHs somewhat implicitly because it's norm is not immediately obvious.

Widely used examples are the Gaussian kernel  $K(x, y) = \exp(-\|x - y\|_2^2/\sigma^2)$ , it produces a space of infinitely smooth functions; and the Laplacian kernel  $K(x, y) = \exp(-\|x - y\|_2/\sigma)$ , it produces the space of Sobolev functions with  $(d + 1)/2$  square integrable derivatives. rkHs spaces are used extensively in numerical analysis [Wen04; FM15] and nonparametric estimation [Wai19; Bac24] due to their analytic and algebraic properties. We use a PSD kernel to form an estimator of the influence function via the spectral representation (2.5). We assume the following properties, and call any such function a Mercer kernel:

ASSUMPTION 3.1 [Mercer kernel]. (0) probability measure  $P$  is absolutely continuous with compact support  $\mathcal{X} \subset \mathbb{R}^d$ ; (i)  $K : \mathcal{X}^2 \rightarrow \mathbb{R}$  is a PSD function; (ii)  $K$  is continuous, bounded, so that its rkHs  $H \subset L^2(P)$  and with the bound  $\kappa := \max_{x \in \mathcal{X}} K(x, x) < +\infty$ ; (iii) for every  $y \in \mathcal{X}$ ,  $\int_{\mathcal{X}} k_y dP = 0$  so that  $H \subset L_0^2(P)$ ; (iv)  $H$  is universal in the sense that it is dense in  $L_0^2(P)$ .

In particular, the Gaussian and Laplacian kernels are universal; see [SFL11] [CS08, ch4.6]. The normalization property (ii) can be imposed on any PSD kernel  $K$  via one step of the Cholesky algorithm  $\tilde{K}(x, y) := K(x, y) - \int_{\mathcal{X}} k_x dP \int_{\mathcal{X}} k_y dP / \int_{\mathcal{X}^2} K dP dP$  is also PSD [PR16, ch4].

#### 3.2 Spectral basis

For a Mercer kernel  $K$ , consider the integral operator  $S_K : L_0^2(P) \rightarrow H \subset L_0^2(P)$  with signature  $K$ :

$$S_K[\phi](x) = \int_{\mathcal{X}} K(x, y) \phi(y) dP(y), \quad \phi \in L_0^2(P). \quad (3.1)$$

Under Assumption 3.1,  $K \in L^2(P \otimes P)$  and  $S$  bounded on  $L_0^2(P)$  by the Cauchy-Schwarz inequality. Operator  $S$  can be thought of as a continuous superposition of slices  $k_y$  of the kernel, and the range of  $S$  is properly contained in  $H$ . Note that the range of  $S$  is the rkHs of the PSD function  $\int k_x k_y dP$ , in particular, it depends on the measure  $P$  [PR16, ch11].

THEOREM 3.2 [Mercer]. Let  $K$  be a Mercer kernel on a compact sample space  $\mathcal{X}$  and  $P$  be a probability measure supported on  $\mathcal{X}$ . Then there is an orthonormal sequence  $\{e_j : \mathcal{X} \rightarrow \mathbb{R}\}_{j \geq 1}$  of continuous  $L_0^2(P)$  eigenfunctions of the integral operator  $S$  with signature  $K$  defined in (3.1), and a corresponding decreasing to zero sequence of eigenvalues, repeated according to multiplicity,  $\{\sigma_j > 0\}_{j \geq 1}$  such that: (i)  $\{e_j\}_{j \geq 1}$  is an orthonormal basis for  $L_0^2(P)$ ; (ii)  $\{\sqrt{\sigma_j} e_j\}_{j \geq 1}$  is an orthonormal basis for  $H$ ; (iii)  $S$  has the diagonalization (2.3), and the adjoint  $S^* : H \rightarrow L_0^2(P)$  is the inclusion operator  $S^* \phi = \phi$ , in particular, relationship (2.4) holds between the  $L_0^2(P)$  and  $H$  inner products.

See [SS09, 4.6], [FM09] for background on integral operators and [SS12; Sun05] for extensions of Mercer's theorem to general, noncompact, domains. In particular, Theorem 3.2 holds for the Gaussian kernel on  $\mathbb{R}^d$  and a measure  $P = \rho \cdot \mathcal{L}^d$  with  $\rho \in L^2(\mathbb{R}^d)$ , as shown in [Sun05, s4].



241 **Example 3.3.** For the Gaussian kernel  $K(x, y) = \exp(-\epsilon^2|x-y|^2)$  on  $\mathbb{R}$  and the centered Gaussian  
 242 weight distribution  $\rho(x) = \alpha \exp(-\alpha^2 x^2)/\sqrt{\pi}$ , the Mercer basis is given by:

$$e_j(x) = \gamma_j e^{-\delta^2 x^2} H_{j-1}(\alpha \beta x), \quad \sigma_j = \sqrt{\frac{\alpha^2}{\alpha^2 + \delta^2 + \epsilon^2}} \left[ \frac{\alpha^2}{\alpha^2 + \delta^2 + \epsilon^2} \right]^{j-1}, \quad j \geq 1, x \in \mathbb{R},$$

243 where  $H_j$  is the Hermite polynomial of degree  $j$ , and constants  $\beta = (1 + [2\epsilon/\alpha]^2)^{1/4}$ ,  $\gamma_j =$   
 244  $(\beta/2^{j-1}\Gamma(j))^{1/2}$ ,  $\delta^2 = \alpha^2(\beta^2 - 1)/2$  are defined in terms of the shape parameters  $\epsilon$  and  $\alpha$ . In  
 245 particular, the eigenvalues  $\sigma_j$  decay exponentially, so that only the first few terms in (2.5) capture  
 246 most of the variation when  $\psi_P$  is smooth. Both  $K$  and  $\rho$  can be extended to  $\mathbb{R}^d$  as tensor products.

### 247 3.3 Nyström method for integral operators

248 In order to estimate the influence function with a given Mercer kernel  $K$  via the spectral representation  
 249 (2.5), the leading eigenvalues  $\{\sigma_j\}_{j=1}^r$  and the corresponding eigenfunctions  $\{e_j\}_{j=1}^r$  of the integral  
 250 operator  $S_K$  are required. If  $P$  is known, these can be computed numerically [FM15, 12.2.2]. If  $P$   
 251 is unknown and only a random sample from  $P$  is available, these must be estimated statistically.  
 252 The Nyström method for approximating the eigendecomposition of the integral operator (3.1) is to  
 253 discretize the integral with the empirical sum, reducing to an eigendecomposition of the empirical  
 254 Gram matrix  $\mathbf{K}_n = [K(X_i, X_j)/n]_{i,j=1}^n$ . This is essentially the well-studied kernel PCA problem to  
 255 estimate the main nonlinear features of  $P$  [Mik+98; ZB05; Sha+05; RBD10; SS22b; SS22a]. A closely related  
 256 problem is the functional PCA [Bos00], and the low-rank Gaussian process approximation [VV+08; BRV19;  
 257 BRV20; SS20]. We find the exposition in [RBD10] particularly lucid and follow it closely.

258 LEMMA 3.4 [Hilbert space LLN]. *Let  $K$  be a Mercer kernel on  $(\mathcal{X}, P)$  generating the rkHs  $H$  and let*  
 259  *$X_1, \dots, X_n$  be an i.i.d. sample from  $P$ . Let the integral operators  $T_H, T_n : H \rightarrow H$  be defined by*

$$T_H[\phi](x) = \int_{\mathcal{X}} \langle \phi, k_y \rangle_H K(x, y) dP(y), \quad T_n[\phi](x) = \frac{1}{n} \sum_{j=1}^n \langle \phi, k_{X_j} \rangle_H K(x, X_j), \quad \text{for } \phi \in H.$$

260 *Then  $T_n \rightarrow T_H$  as  $n \rightarrow \infty$  in the Hilbert-Schmidt norm in probability, and, for each  $n \geq 1$ ,*

$$\|T_H - T_n\|_{\text{HS}} \lesssim \frac{\kappa \sqrt{\tau}}{\sqrt{n}} \quad (3.2)$$

261 *with probability at least  $1 - 2e^{-\tau}$ .*

262 Note that  $T_n$  is the empirical analogue of  $T_H$  obtained by replacing the continuous integral with  
 263 respect to  $P$  by the discrete integral with respect to the empirical distribution of a random sample  
 264 from  $P$ . By appealing to a suitable law of large numbers or concentration inequality, it follows that  
 265  $T_n$  is consistent for  $T_H$ . Furthermore,  $T_H$  is related to  $S_K$ , whereas  $T_n$  is related to  $\mathbf{K}_n$ , providing a  
 266 link between the continuous operator  $S_K$  and the matrix  $\mathbf{K}_n$  that is otherwise not immediately clear.

267 Recall that  $S^*$  is the inclusion of  $H$  into  $L_0^2(P)$ , and note that the operators  $T_H = SS^*$  and  
 268  $T_K := S^*S$  are essentially the same operator with the same action on all functions  $\phi \in H$ , differing  
 269 only in the domain of definition and the embedding space of the range. In particular, the eigenvalues  
 270 of  $T_K$ ,  $S$  and  $T_H$  are exactly the same and the eigenfunctions of  $T_K$  and  $T_H$  are related by the  
 271 inclusion (resp. regularization) operators  $S^*$  (resp.  $S$ ), in other words are the same functions but  
 272 viewed as elements of  $L_0^2(P)$  and  $H$  respectively.

273 Furthermore, the finite-rank empirical operator  $T_n$  and the empirical Gram matrix  $\mathbf{K}_n$ , are similarly  
 274 related via the Nyström restriction operator  $R_n : H \rightarrow \mathbb{R}^n$  given by  $R_n[\phi] = (\phi(X_j))_{j=1}^n$ . Its  
 275 adjoint  $R_n^* : \mathbb{R}^n \rightarrow H$  is the Nyström extension operator given by  $R_n^*[\mathbf{y}] = \sum_{j=1}^n y^j k_{X_j}/n$   
 276 for  $\mathbf{y} = (y^1, \dots, y^n) \in \mathbb{R}^n$  endowed with inner product  $\langle \mathbf{x}, \mathbf{y} \rangle_n := \sum_{j=1}^n x^j y^j / n$ . With this  
 277 notation, we have  $T_n = R_n^* R_n$  and  $\mathbf{K}_n = R_n R_n^*$ , from which it follows that  $T_n$  and  $\mathbf{K}_n$  have the  
 278 same nonzero eigenvalues and the corresponding eigenvectors are related via the restriction (resp.  
 279 extension) operators  $R_n$  (resp.  $R_n^*$ ).

280 The next result is an application of perturbation bound of [Kat87], see also [RBD10], to infer consistency  
 281 of the spectrum of  $\mathbf{K}_n$  for the spectrum of  $T_K = S^*S$ .

LEMMA 3.5 [Consistency of eigenvalues]. Let  $K$  be a Mercer kernel on  $(\mathcal{X}, P)$  generating the rkHs  $H$  and let  $X_1, \dots, X_n$  be an i.i.d. sample from  $P$ . Let the integral operator  $T_K : L_0^2(P) \rightarrow L_0^2(P)$  and the empirical Gram matrix multiplication operator  $\mathbf{K}_n : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be given by

$$T_K[\phi](x) = \int_{\mathcal{X}} K(x, y) \phi(y) dP(y), \quad \mathbf{K}_n[\mathbf{y}] = \left[ \frac{1}{n} K(X_i, X_j) \right]_{i,j=1}^n \mathbf{y}, \quad \phi \in L_0^2(P), \quad \mathbf{y} \in \mathbb{R}^n.$$

Let  $\{\sigma_j\}_{j \geq 1}$  be the decreasing enumeration of the eigenvalues of  $T_K$ , repeated according to the multiplicity, and let  $\{\hat{\sigma}_j\}_{j \geq 1}$  denote the analogous enumeration of the eigenvalues of  $\mathbf{K}_n$ , extended by zero. Then  $\hat{\sigma}_j \rightarrow \sigma_j$  as  $n \rightarrow \infty$  uniformly in probability, and, for each  $n \geq 1$ ,

$$\sup_{j \geq 1} |\sigma_j - \hat{\sigma}_j| \leq \|T_K - \mathbf{K}_n\|_{\text{HS}} \lesssim \frac{\kappa \sqrt{\tau}}{\sqrt{n}}, \quad (3.3)$$

and  $\sum_{j \geq 1} (\sigma_j - \hat{\sigma}_j)^2 \leq \|T_K - \mathbf{K}_n\|_{\text{HS}}^2 \lesssim \frac{\kappa^2 \tau}{n}$  and  $|\sum_{j \geq 1} (\sigma_j - \hat{\sigma}_j)| = |\text{tr}(T_H) - \text{tr}(T_n)| \lesssim \frac{\kappa \sqrt{\tau}}{\sqrt{n}}$  with probability at least  $1 - 2e^{-\tau}$ .

For  $N \in \mathbb{N}$ , let  $r(N)$  denote the total number of eigenvalues, accounting for multiplicity, corresponding to the leading  $N$  distinct eigenvalues  $\sigma_{r(1)} > \dots > \sigma_{r(N)}$  and let  $\sigma_{r(N)+1}$  be the next largest distinct eigenvalue of  $T_K$ . Let  $H_N := \text{span}\{\mathbf{e}_1, \dots, \mathbf{e}_{r(N)}\} \subset H \subset L_0^2(P)$  be the eigenspace of the  $N$  leading distinct eigenvalues, and let

$$P_N : H \rightarrow H_N, \quad P_N[\phi] := \sum_{j=1}^{r(N)} \langle \phi, \sqrt{\sigma_j} \mathbf{e}_j \rangle_H \sqrt{\sigma_j} \mathbf{e}_j = \sum_{j=1}^{r(N)} \langle \phi, \mathbf{e}_j \rangle_{L^2(P)} \mathbf{e}_j, \quad \phi \in H$$

be its spectral projection. The following perturbation bound was used by [ZB05; KG00]: if  $\hat{T}_K$  is a finite-rank estimate of the operator  $T_K$  with precision on the order of the  $N$ th spectral gap with  $\|T_K - \hat{T}_K\|_{\text{op}} \leq [\sigma_{r(N)} - \sigma_{r(N)+1}]/4$ , then the eigenspaces  $H_N$  of  $T_K$  and  $\hat{H}_N$  of the leading  $r(N)$  eigenvalues of  $\hat{T}_K$  must also be close with  $\|P_N - P_{\hat{H}_N}\|_{\text{op}} \leq 2/[\sigma_{r(N)} - \sigma_{r(N)+1}] \|T_K - \hat{T}_K\|_{\text{op}}$ .

LEMMA 3.6 [Consistency of spectral projections]. In the setting of Lemma 3.5, let  $\mathbf{y}_1, \dots, \mathbf{y}_n$  denote the orthonormal (for the scalar product  $\langle \cdot, \cdot \rangle_n = \langle \cdot, \cdot \rangle_{\mathbb{R}^n}/n$  so that  $\|\mathbf{y}_i\|_{\mathbb{R}^n} = \sqrt{n}$ ) eigenvectors of the empirical Gram matrix  $\mathbf{K}_n$ :

$$\mathbf{K}_n[\mathbf{y}] = \sum_{i=1}^n \hat{\sigma}_i \langle \mathbf{y}, \mathbf{y}_i \rangle_n \mathbf{y}_i, \quad \mathbf{y} \in \mathbb{R}^n. \quad (3.4)$$

Then, with  $\mathbf{y}_i = (y_i^1, \dots, y_i^n) \in \mathbb{R}^n$ ,

$$\hat{\mathbf{e}}_i(x) = \frac{1}{\hat{\sigma}_i} R_n^*[\mathbf{y}_i](x) = \frac{1}{\hat{\sigma}_i} \frac{1}{n} \sum_{j=1}^n y_i^j K(x, X_j), \quad i = 1, \dots, n \quad (3.5)$$

denote the eigenfunctions of the empirical integral operator  $T_n$  (with normalization  $\|\sqrt{\hat{\sigma}_i} \hat{\mathbf{e}}_i\|_H \equiv 1$  and  $\|\hat{\mathbf{e}}_i\|_{L^2(P)} \approx 1$ ):

$$T_n[\phi](x) = \sum_{i=1}^n \hat{\sigma}_i \langle \phi, \sqrt{\hat{\sigma}_i} \hat{\mathbf{e}}_i \rangle_H \sqrt{\hat{\sigma}_i} \hat{\mathbf{e}}_i(x), \quad \phi \in H.$$

Let  $N \in \mathbb{N}$  and  $P_N$  (resp.  $P_{\hat{H}_N}$ ) denote the spectral projection operator on the eigenspace of the leading  $r(N)$  eigenvalues of  $T_K$  (resp.  $T_n$ ) and  $I_H$  denote the identity on  $H$ . Then, for any  $\psi \in H$ ,

$$P_{\hat{H}_N}[\psi] = \sum_{j=1}^{r(N)} \hat{\sigma}_j \langle \psi, \hat{\mathbf{e}}_j \rangle_H \hat{\mathbf{e}}_j \rightarrow P_N[\psi] = \sum_{j=1}^{r(N)} \sigma_j \langle \psi, \mathbf{e}_j \rangle_H \mathbf{e}_j, \quad \text{as } n \rightarrow \infty$$

in  $H$  in probability, and, if  $n \geq 128\kappa^2\tau/[\sigma_N - \sigma_{N+1}]^2$ , then

$$\sum_{j=1}^{r(N)} \|(I_H - P_N) \sqrt{\hat{\sigma}_j} \hat{\mathbf{e}}_j\|_H^2 + \sum_{j=r(N)+1}^n \|P_N \sqrt{\hat{\sigma}_j} \hat{\mathbf{e}}_j\|_H^2 \leq \frac{32\kappa^2\tau}{(\sigma_{r(N)} - \sigma_{r(N)+1})^2 n}$$

with probability at least  $1 - 2e^{-\tau}$ .



### 3.4 Kernel von Mises estimator

**THEOREM 3.7** [consistency of the kernel von Mises estimator]. *Let  $K$  be a Mercer kernel on  $(\mathcal{X}, P)$  generating the rkHs  $H$  and let  $X_1, \dots, X_n$  be an i.i.d. sample from  $P$ . Let  $\theta$  be a pathwise differentiable functional on  $\mathcal{P}$  with derivative  $D\theta_P$  and influence function  $\psi_P \in L_0^2(P)$  for  $P \in \mathcal{P}$ . Assume that  $D\theta_P$ , equivalently  $\psi_P$ , are continuous in  $P \in \mathcal{P}$  in an appropriate sense and that  $\hat{f}$  is a consistent estimator of the density of  $P$  in a compatible notion of convergence. Let  $\{\hat{e}_j\}$  and  $\{\hat{\sigma}_j\}$  be the Nyström estimators of the eigenfunctions and eigenvalues of  $T_K$ . For a fixed rank  $1 \leq r \leq n$  and regularization loading  $\lambda \geq 0$ , let*

$$\psi_\lambda^r(x) := \sum_{j=1}^r \frac{D\theta_P[e_j]}{1 + 2\lambda/\sigma_j} e_j(x), \quad \hat{\psi}_\lambda^r(x) := \sum_{j=1}^r \frac{1}{1 + 2\lambda/\hat{\sigma}_j} \left[ \frac{d}{dt} \theta(\hat{f}_t^j) \right]_{t=0} \hat{e}_j(x) \quad (3.6)$$

*denote the rank- $r$  approximation of  $\psi_P$  and its plug-in estimator obtained by replacing the unknown  $\sigma_j, e_j, D\theta_P[e_j]$  with their estimates. The pathwise derivative can be computed along, e.g., the linear perturbation  $\hat{f}_t^j = [1 + t\hat{e}_j]\hat{f}$  since the scores  $e_j$  and  $\hat{e}_j$  are bonded. Then  $\|\hat{\psi}_\lambda^r - \psi_{P,\lambda}^r\|_H \rightarrow 0$  as  $n \rightarrow \infty$  in probability and there exist sequences  $r(n) \rightarrow \infty$  increasing and  $\lambda(n) \rightarrow 0$  decreasing such that  $\|\hat{\psi}_\lambda^r - \psi_P\|_{L^2(P)} \rightarrow 0$  as  $n \rightarrow \infty$  in probability.*

**Limitations and future work.** Our spectral formula (2.5) and kernel implementation (3.6) aim to enable automation of methods based on asymptotic analysis and first-order techniques. It is important to point out that these methods, whether carried out analytically or numerically, require theoretical justification and provide approximations that might be more or less accurate in any particular problem. Applications that require our estimator, analogously to the bootstrap methods for statistical inference [Efr92], require theoretical justification, which limits the utility it provides. Nonetheless, it is hoped that the theoretical results provided here will allow to leverage efficient computation with PSD kernels [RCR15; Ste+20; Che+25], not considered here, for the applications describe in the Introduction. Further theoretical work needs to be done: Downstream tasks, such as debaised machine learning [Che+18], make assumptions on the rate of convergence, e.g.,  $o(n^{-1/4})$ , of the influence function estimator. While we discuss some of the ingredients to study these rates for our estimator, we only show consistency here and will investigate the rates in follow-up work.

**Simulation experiments.** As a toy experiment, we compute the oracle low-rank regularization  $\psi_\lambda^r$  and its estimator  $\hat{\psi}_\lambda^r$  as well as the distribution of the estimation error  $\|\psi_\lambda^r - \hat{\psi}_\lambda^r\|_{L^2(P)}$  for the mean functional  $\theta = E[X]$  in the setting of Example 3.3.

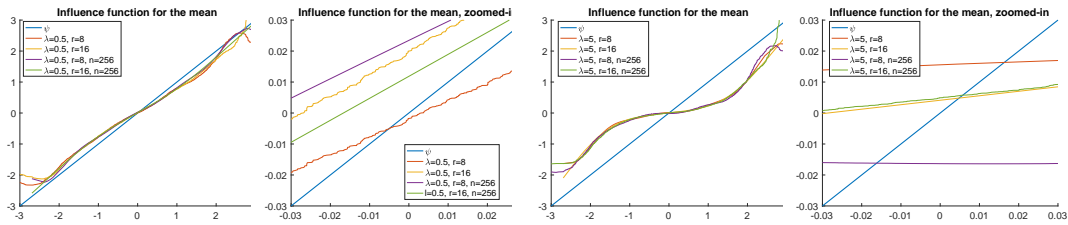


Figure 1: Influence function  $\psi$ , oracle low-rank regularization  $\psi_\lambda^r$  and estimator  $\hat{\psi}_{\lambda,n}^r$ .

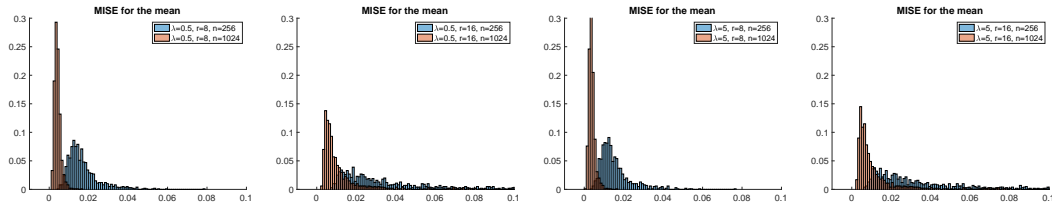


Figure 2: Sampling distribution of the error  $\|\psi_\lambda^r - \hat{\psi}_{\lambda,n}^r\|_{L^2(P)}$  based on  $10^3$  Monte Carlo experiments.

## References

- [AGS17] I. Andrews, M. Gentzkow, and J. M. Shapiro. “Measuring the Sensitivity of Parameter Estimates to Estimation Moments”. *The Quarterly Journal of Economics* (2017).
- [AGS20a] I. Andrews, M. Gentzkow, and J. M. Shapiro. “On the informativeness of descriptive statistics for structural estimates”. *Econometrica* 88.6 (2020), pp. 2231–2258.
- [AGS20b] I. Andrews, M. Gentzkow, and J. M. Shapiro. “Transparency in structural research”. *Journal of Business & Economic Statistics* 38.4 (2020), pp. 711–722.
- [Bac24] F. Bach. *Learning theory from first principles*. MIT press, 2024.
- [BGM20] T. Broderick, R. Giordano, and R. Meager. “An automatic finite-sample robustness metric: when can dropping a little data make a big difference?”. *arXiv preprint arXiv:2011.14999* (2020).
- [Bic+93] P. J. Bickel et al. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.
- [Bic82] P. J. Bickel. “On Adaptive Estimation”. *Ann. Statist.* 10.3 (Sept. 1982), pp. 647–671.
- [Bos00] D. Bosq. *Linear processes in function spaces: theory and applications*. Vol. 149. Springer Science & Business Media, 2000.
- [BR07] V. I. Bogachev and M. A. S. Ruas. *Measure theory*. Vol. 1. 1. Springer, 2007.
- [BRV19] D. Burt, C. E. Rasmussen, and M. Van Der Wilk. “Rates of convergence for sparse variational Gaussian process regression”. *International Conference on Machine Learning*. PMLR, 2019, pp. 862–871.
- [BRV20] D. R. Burt, C. E. Rasmussen, and M. Van Der Wilk. “Convergence of sparse variational inference in Gaussian processes regression”. *Journal of Machine Learning Research* 21.131 (2020), pp. 1–63.
- [CC23] T. Christensen and B. Connault. “Counterfactual sensitivity and robustness”. *Econometrica* 91.1 (2023), pp. 263–298.
- [Cha86] G. Chamberlain. “Asymptotic efficiency in semi-parametric models with censoring”. *journal of Econometrics* 32.2 (1986), pp. 189–218.
- [Che+18] V. Chernozhukov et al. *Double/debiased machine learning for treatment and structural parameters*. 2018.
- [Che+22] V. Chernozhukov et al. “Locally robust semiparametric estimation”. *Econometrica* 90.4 (2022), pp. 1501–1535.
- [Che+25] Y. Chen et al. “Randomly pivoted Cholesky: Practical approximation of a kernel matrix with few entry evaluations”. *Communications on Pure and Applied Mathematics* 78.5 (2025), pp. 995–1041.
- [Cho+24] B. Cho et al. “Kernel debiased plug-in estimation: Simultaneous, automated debiasing without influence functions for many target parameters”. *Proceedings of machine learning research* 235 (2024), p. 8534.
- [CLV19] M. Carone, A. R. Luedtke, and M. J. Van Der Laan. “Toward computerized efficient estimation in infinite-dimensional models”. *Journal of the American Statistical Association* (2019).
- [CS08] A. Christmann and I. Steinwart. “Support vector machines” (2008).
- [CS18] X. Chen and A. Santos. “Overidentification in regular models”. *Econometrica* 86.5 (2018), pp. 1771–1817.
- [Dud18] R. M. Dudley. *Real analysis and probability*. Chapman and Hall/CRC, 2018.
- [Efr92] B. Efron. “Bootstrap methods: another look at the jackknife”. *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 569–593.
- [FM09] J. Ferreira and V. Menegatto. “Eigenvalues of integral operators defined by smooth positive definite kernels”. *Integral Equations and Operator Theory* 64.1 (2009), pp. 61–81.
- [FM15] G. E. Fasshauer and M. J. McCourt. *Kernel-based approximation methods using Matlab*. Vol. 19. World Scientific Publishing Company, 2015.
- [FS19] Z. Fang and A. Santos. “Inference on directionally differentiable functions”. *The Review of Economic Studies* 86.1 (2019), pp. 377–412.
- [Gro+23] R. Grosse et al. “Studying large language model generalization with influence functions”. *arXiv preprint arXiv:2308.03296* (2023).
- [Hin+22] O. Hines et al. “Demystifying statistical learning based on efficient influence functions”. *The American Statistician* 76.3 (2022), pp. 292–304.
- [HM95] J. L. Horowitz and C. F. Manski. “Identification and robustness with contaminated and corrupted data”. *Econometrica: Journal of the Econometric Society* (1995), pp. 281–302.
- [Hub72] P. J. Huber. “The 1972 wald lecture robust statistics: A review”. *The Annals of Mathematical Statistics* 43.4 (1972), pp. 1041–1067.
- [Hub92] P. J. Huber. “Robust estimation of a location parameter”. *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

- [IN22] H. Ichimura and W. K. Newey. “The influence function of semiparametric estimators”. *Quantitative Economics* 13.1 (2022), pp. 29–61.
- [JWZ22] M. Jordan, Y. Wang, and A. Zhou. “Empirical gateaux derivatives for causal inference”. *Advances in Neural Information Processing Systems* 35 (2022), pp. 8512–8525.
- [Kat87] T. Kato. “Variation of discrete spectra”. *Communications in Mathematical Physics* 111 (1987), pp. 501–504.
- [Ken24] E. H. Kennedy. “Semiparametric doubly robust targeted double machine learning: a review”. *Handbook of Statistical Methods for Precision Medicine* (2024), pp. 207–236.
- [KG00] V. Koltchinskii and E. Giné. “Random matrix approximation of spectra of integral operators” (2000).
- [KL17] P. W. Koh and P. Liang. “Understanding black-box predictions via influence functions”. *International conference on machine learning*. PMLR. 2017, pp. 1885–1894.
- [KL76] Y. A. Koshevnik and B. Y. Levit. “On a Non-Parametric Analogue of the Information Matrix”. *Theory of Probability & Its Applications* 21.4 (1976), pp. 738–753.
- [Kla87] C. A. Klaassen. “Consistent estimation of the influence function of locally asymptotically linear estimators”. *The Annals of Statistics* 15.4 (1987), pp. 1548–1562.
- [Lue97] D. G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.
- [Mad+17] A. Madry et al. “Towards deep learning models resistant to adversarial attacks”. *arXiv preprint arXiv:1706.06083* (2017).
- [Mik+98] S. Mika et al. “Kernel PCA and de-noising in feature spaces”. *Advances in neural information processing systems* 11 (1998).
- [Mis47] R. von Mises. “On the asymptotic distribution of differentiable statistical functions”. *The annals of mathematical statistics* 18.3 (1947), pp. 309–348.
- [Muk18] Y. Mukhin. “Sensitivity of regular estimators”. *arXiv preprint arXiv:1805.08883* (2018).
- [Muk19] Y. Mukhin. “Geometric methods in econometrics and statistics”. PhD thesis. Massachusetts Institute of Technology, 2019.
- [Muk21] Y. Mukhin. *On Robustness of Counterfactuals in Structural Models*. Presented at the NeurIPS Workshop on Robustness and misspecification in probabilistic modeling. 2021.
- [New94] W. K. Newey. “The asymptotic variance of semiparametric estimators”. *Econometrica: Journal of the Econometric Society* (1994), pp. 1349–1382.
- [Pin12] I. Pinelis. “Optimum bounds for the distributions of martingales in Banach spaces”. *arXiv preprint arXiv:1208.2200* (2012).
- [PR16] V. I. Paulsen and M. Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*. Vol. 152. Cambridge university press, 2016.
- [Pru+20] G. Pruthi et al. “Estimating training data influence by tracing gradient descent”. *Advances in Neural Information Processing Systems* 33 (2020), pp. 19920–19930.
- [RBD10] L. Rosasco, M. Belkin, and E. De Vito. “On Learning with Integral Operators.” *Journal of Machine Learning Research* 11.2 (2010).
- [RCR15] A. Rudi, R. Camoriano, and L. Rosasco. “Less is more: Nyström computational regularization”. *Advances in neural information processing systems* 28 (2015).
- [Roy10] H. Royden. *Real Analysis*. 2010.
- [Sem20] V. Semenova. “Generalized lee bounds”. *arXiv preprint arXiv:2008.12720* (2020).
- [Sem25] V. Semenova. “Debiased Machine Learning of Aggregated Intersection Bounds and Other Causal Parameters”. *Available at SSRN 5134514* (2025).
- [SFL11] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. “Universality, Characteristic Kernels and RKHS Embedding of Measures.” *Journal of Machine Learning Research* 12.7 (2011).
- [Sha+05] J. Shawe-Taylor et al. “On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA”. *IEEE Transactions on Information Theory* 51.7 (2005), pp. 2510–2522.
- [SS09] E. M. Stein and R. Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.
- [SS12] I. Steinwart and C. Scovel. “Mercer’s theorem on general domains: On the interaction between measures, kernels, and RKHSs”. *Constructive Approximation* 35 (2012), pp. 363–417.
- [SS20] A. Solin and S. Särkkä. “Hilbert space methods for reduced-rank Gaussian process regression”. *Statistics and Computing* 30.2 (2020), pp. 419–446.
- [SS22a] B. K. Sriperumbudur and N. Sterge. “Approximate kernel PCA: Computational versus statistical trade-off”. *The Annals of Statistics* 50.5 (2022), pp. 2713–2736.
- [SS22b] N. Sterge and B. K. Sriperumbudur. “Statistical optimality and computational efficiency of nystrom kernel pca”. *Journal of Machine Learning Research* 23.337 (2022), pp. 1–32.

- [Ste+20] N. Sterge et al. “Gain with no pain: Efficiency of kernel-pca by nyström sampling”. *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3642–3652.
- [Sun05] H. Sun. “Mercer theorem for RKHS on noncompact sets”. *Journal of Complexity* 21.3 (2005), pp. 337–349.
- [Vaa00] A. W. van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.
- [Vaa91] A. W. van der Vaart. “On differentiable functionals”. *The Annals of Statistics* (1991), pp. 178–204.
- [Vil03] C. Villani. *Topics in optimal transportation*. Vol. 58. American Mathematical Soc., 2003.
- [VR06] M. J. Van Der Laan and D. Rubin. “Targeted maximum likelihood learning”. *The international journal of biostatistics* 2.1 (2006).
- [VV+08] A. W. Van Der Vaart, J. H. Van Zanten, et al. “Reproducing kernel Hilbert spaces of Gaussian priors”. *IMS Collections* 3 (2008), pp. 200–222.
- [Wai19] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.
- [Wen04] H. Wendland. *Scattered data approximation*. Vol. 17. Cambridge university press, 2004.
- [ZB05] L. Zwald and G. Blanchard. “On the convergence of eigenspaces in kernel principal component analysis”. *Advances in neural information processing systems* 18 (2005).

## A Pathwise derivatives and von Mises formula

**Tangent space** [KL76], [Bic+93, s3.2], [Vaa00, s25.3] and also [Vil03, s8.1.2] At each  $P \in \mathcal{P}$  we consider perturbations to  $P$  in  $\mathcal{P}$  along one-dimensional parametric submodels  $t \mapsto P_t \in \mathcal{P}$  with parameter  $t \in [0, \epsilon)$  and  $P = P_{t=0}$ . These perturbations must be smooth and admit infinitesimal directions of perturbations; if we think of  $\{P_t\}_t$  as a curve through  $P$  in the space of probability measures, what is required is that it has a tangent vector at  $P$ . Let’s assume that measures in our model are absolutely continuous  $P = \rho \cdot \mathcal{L}^d$  with density function  $\rho$  with respect to the Lebesgue measure  $\mathcal{L}^d$  on  $\mathbb{R}^d$ . The direction of perturbation  $P_t$  can then be identified with the time-derivative of the density along the curve  $\partial_t \rho_t(x)$  for each point  $x \in \mathcal{X} \subset \mathbb{R}^d$ . The kinds of perturbations that are relevant for defining the influence function do not change the support of the distribution  $P$ , so  $\partial_t \rho_t / \rho$  is well-defined, and it turns out to be more convenient mathematically to work with the *score function*

$$\phi(x) := \partial_t|_{t=0} \log \rho_t(x), \quad x \in \mathcal{X} \subset \mathbb{R}^d. \quad (\text{A.1})$$

The score function  $\phi : \mathcal{X} \rightarrow \mathbb{R}^2$  is the tangent vector to the curve  $P_t$ , the infinitesimal change in  $P$  along the curve. The derivative in (A.1), need not hold pointwise, but rather in the Hellinger norm:

$$\lim_{t \rightarrow 0} \int_{\mathcal{X}} \left[ t^{-1}(\sqrt{\rho_t} - \sqrt{\rho}) - 2^{-1}\phi(x)\sqrt{\rho} \right]^2 d\mathcal{L}^d(x). \quad (\text{A.2})$$

The existence of this limit implies that  $\int \phi dP = 0$  and  $\int \phi^2 dP < +\infty$ . We denote the space of all such function by  $L_0^2(P)$ , indicating with the subscript that it is the subspace of  $L^2(P)$  of all functions that have  $P$ -mean zero.

**Pathwise derivative** A functional  $\theta : \mathcal{P} \rightarrow \mathbb{R}$  is pathwise differentiable at  $P$  (with respect to a collection of paths) if, (i) for a given regular path  $P_t$  the composition  $t \mapsto \theta(P_t)$  is a differentiable function from  $[0, \epsilon)$  to  $\mathbb{R}$  at time  $t = 0$ ; and (ii) there is a bounded linear map  $D\theta_P : L_0^2(P) \rightarrow \mathbb{R}$  such that

$$\lim_{t \rightarrow 0} t^{-1}[\theta(P_t) - \theta(P)] = D\theta_P[\phi] \quad (\text{A.3})$$

for every score function  $\phi \in L_0^2(P)$  and every admissible path  $P_t$  with score  $\phi$ . The definitions of the score and pathwise derivative are those of Riemannian geometry that extends techniques of calculus to nonlinear spaces. Notions of smoothness are more nuanced in this infinite dimensional setting, see the penultimate paragraph in [Vil03, s3.2.3]

**Influence function** By Reisz’ representation theorem for Hilbert spaces [SS09, 4.5], [Dud18, 5.5.1], for the bounded linear functional  $D\theta_P : L_0^2(P) \rightarrow \mathbb{R}$  there exists a fixed score  $\psi_P \in L_0^2(P)$  such that the action of the derivative  $D\theta_P$  on any score  $v$  has the following representation in terms of the inner product:

$$D\theta_P[\phi] = \langle \phi, \psi_P \rangle_{L^2(P)}, \quad \text{for all } \phi \in L_0^2(P). \quad (\text{A.4})$$

497 The Riesz representer  $\psi_P$  of the derivative functional of parameter  $\theta(P)$  is known as the influence  
 498 function. It has the useful geometric interpretation of the gradient score for parameter  $\theta$ , i.e., the  
 499 direction of perturbation to  $P$  such that the functional changes most rapidly:

$$D\theta_P[\phi] = \langle \phi, \psi_P \rangle_{L^2(P)} \leq \|\phi\|_{L^2(P)} \|\psi_P\|_{L^2(P)}, \quad \phi \in L^2_0(P) \quad (\text{A.5})$$

500 where equality holds if and only if  $v = c\psi$  by the Cauchy-Schwarz inequality. If we restrict the  
 501 norm of the perturbation  $\|\phi\|_{L^2(P)} \leq 1$  as in Lemma 2.1 and use linearity of  $D\theta_P$ , it follows that  
 502 the unique maximum in (A.5) is achieved at the score  $\psi_P/\|\psi_P\|_{L^2(P)}$  and the norm of the influence  
 503 function  $\|\psi_P\|_{L^2(P)}$  is the largest sensitivity of  $\theta$  to a perturbation at  $P$ .

#### 504 A.1 Proof of Theorem 1.1

505 We begin with a score calculation for the original von Mises [Mis47] calculation with a point mass  
 506 perturbation. The example shows that this singular perturbation does not have a score function, hence  
 507 the need to smooth out the point mass in general applications of this technique as in [IN22; CLV19].

508 **Example A.1 Score for the von Mises calculation.** Let  $P$  and  $\delta_x$  be a continuous distribution and  
 509 a point mass on  $\mathcal{X}$ . Take the path  $[0, 1] \ni t \mapsto P_t = (1-t)P_t + t\delta_x$  for the calculations of von  
 510 Mises and Huber [Hub72; Hub92]:

$$\partial_{t|t=0}\theta(P_t) = \lim_{t \rightarrow 0} t^{-1}[\theta(P_t) - \theta(P)] = \int_{\mathcal{X}} \psi_P d[\delta_x - P] = \psi_P(x) \quad (\text{A.6})$$

511 which can be made rigorous if  $\psi_P$  is, for instance, continuous at  $x$ . However, this perturbation  
 512 to  $P$  is not smooth in the sense of differentiability in quadratic mean (A.2). We compute the  
 513 tangent vector to  $P_t$  at  $t = 1/3$  and  $t = 0$ . Take  $\mu = P + \delta_x$  to be the dominating measure for  
 514 the path, so that  $f_t(z) = (1-t)\mathbb{1}_{\mathcal{X} \setminus x}(z) + t\mathbb{1}_{\{x\}}(z)$  is the Radon-Nikodym derivative at time  
 515  $t$ . The corresponding embedding of  $P_t$  into the space of square roots of measures  $H_2$  [BR07, c4] is  
 516  $\sqrt{f}_t(z) = \sqrt{1-t}\mathbb{1}_{\mathcal{X} \setminus x}(z) + \sqrt{t}\mathbb{1}_x(z)$ . Note that the path is no longer linear in the embedding  
 517 space. Also note that  $d_{\text{TV}}(P_t, P_{t+h}) = 2 \sup_A |P_{t+h}[A] - P_t[A]| = 2h$  is continuous in  $H_2$ . For  
 518  $t = 1/3$ , the density  $\sqrt{f}_t(z)$  can be differentiated pointwise for each  $z \in \mathcal{X}$  to find the score function  
 519  $\frac{1}{2}\phi_{\frac{1}{3}}(z)\sqrt{f}_{\frac{1}{3}}(z) = -\frac{1}{2}[\frac{2}{3}]^{-1/2}\mathbb{1}_{\mathcal{X} \setminus x}(z) + \frac{1}{2}[\frac{1}{3}]^{-1/2}\mathbb{1}_x(z)$  and verify differentiability in quadratic  
 520 mean

$$\begin{aligned} & \left\| t^{-1}[\sqrt{f}_{\frac{1}{3}+t} - \sqrt{f}_{\frac{1}{3}}] - \frac{1}{2}\phi_{\frac{1}{3}}(z)\sqrt{f}_{\frac{1}{3}}(z) \right\|_{H_2}^2 \\ &= \left\{ t^{-1} \left[ \sqrt{\frac{2}{3}-t} - \sqrt{\frac{2}{3}} \right] - (-1)\frac{1}{2}(\frac{2}{3})^{-1/2} \right\}^2 P[\mathcal{X} \setminus x] \\ & \quad + \left\{ t^{-1} \left[ \sqrt{\frac{1}{3}+t} - \sqrt{\frac{1}{3}} \right] - \frac{1}{2}(\frac{1}{3})^{-1/2} \right\}^2 \delta_{\{x\}}[x] \\ &= o(1) \quad \text{as } t \rightarrow 0. \end{aligned}$$

521 Repeating the calculation with  $t = 0$ , we note that the right derivative of  $\sqrt{t}$  is infinite, so there is  
 522 no score function with finite  $\mu$ -a.e. values that can satisfy (A.2). Consequently, the path  $P_t$  is not  
 523 smooth in the Hellinger norm and does not have a tangent vector at  $t = 0$ .

524 To remedy the lack of smoothness and extend the von Mises formula (A.6) to all pathwise differen-  
 525 tiable functionals, the point mass perturbations must be mollified.

526 **LEMMA A.2** [Approximation to von Mises perturbation with a score]. *Suppose  $K$  is a bounded probability*  
 527 *density function on  $\mathbb{R}^d$  with support in the unit ball  $|x| \leq 1$ . Then*

$$K^\delta(x) := \delta^{-d}K(\delta^{-1}x), \quad \delta > 0 \quad (\text{A.7})$$

528 *is an approximation to the identity in the sense of [SS09, p109], that is*

- 529 (i)  $\int_{\mathbb{R}^d} K^\delta(x) dx = 1$ .
- 530 (ii)  $|K^\delta(x)| \leq A\delta^{-d}$  for all  $\delta > 0$ .
- 531 (iii)  $|K^\delta(x)| \leq A\delta/|x|^{d+1}$  for all  $\delta > 0$  and  $x \in \mathbb{R}^d$ .

532 Here  $A$  is a constant independent of  $\delta$ .

533 Suppose  $P_0$  is a probability measure that is absolutely continuous with respect to the Lebesgue  
534 measure  $\mathcal{L}^d$  with a continuous density function  $f_0$ . Let

$$K^{\delta,z}(x) := \left[ \int_{\{f_0 > \delta\}} K^\delta(z-x) dx \right]^{-1} \mathbb{1}_{\{f_0 > \delta\}}(x) K^\delta(z-x), \quad (\text{A.8})$$

535 then for  $z \in \{f_0 > 0\}$  we have  $K^{\delta,z}(x) = K^\delta(z-x)$  for all sufficiently small  $\delta > 0$  (which depend  
536 on  $z$  that is fixed throughout). Furthermore,

$$f_t^{\delta,z}(x) := (1-t)f_0(x) + tK^{\delta,z}(x) \quad (\text{A.9})$$

537 is a curve of probability densities with parameter  $t$  in an interval around 0, that is differentiable in  
538 quadratic mean (A.2) at  $t = 0$  with the score function

$$\phi_{\delta,z}(x) := \frac{d}{dt} \log f_t^{\delta,z}(x) = \frac{K^{\delta,z}(x)}{f_0(x)} - 1. \quad (\text{A.10})$$

539

540 *Proof.* The three properties of an approximation to the identity follow respectively from dilation  
541 invariance of Lebesgue integral, boundedness and compact support of the kernel  $K$ .

542 Fix a  $z \in \{f_0 > 0\}$ . By the continuity of  $f_0$  there is a neighborhood  $\mathcal{N}$  of  $z$  such that  $f_0$  is bounded  
543 away from zero on  $\mathcal{N}$ . For all  $\delta > 0$  small enough,  $x \mapsto K^\delta(z-x)$  is supported in  $\mathcal{N}$  by bounded  
544 support and dilation construction, so that  $K^\delta(z-x) \equiv K^{\delta,z}(x)$ . Therefore for  $t$  negative and close  
545 enough to 0, function  $f_{t,\delta,z}$  is a well-defined probability density and its score functions

$$\phi_{t,\delta,z}(x) := \frac{d}{dt} \log f_t^{\delta,z}(x) = \frac{K^{\delta,z}(x) - f_0(x)}{f_t^{\delta,z}(x)} \quad (\text{A.11})$$

546 are bounded in  $x \in \mathcal{X}$ . To check (A.2)

$$\int_{\mathcal{X}} \left[ \frac{\sqrt{f_{t,\delta,z}} - \sqrt{f_0}}{t} - \frac{1}{2} \phi_{\delta,z} \sqrt{f_0} \right]^2 dx \rightarrow 0 \quad \text{as } t \rightarrow 0, \quad (\text{A.12})$$

547 note that the map  $t \mapsto \sqrt{f_{t,\delta,z}}(x)$  is continuously differentiable for each  $x$  in a neighborhood  
548  $t \in (-\epsilon, \epsilon)$  of 0 with the derivative  $\frac{1}{2} v_{t,\delta,z}(x) \sqrt{f_{t,\delta,z}}(x)$ , therefore the problem is to justify the  
549 change of order of the limit  $t \rightarrow 0$  and the integral  $\int_{\mathcal{X}} dx$  in (A.12). By the fundamental theorem of  
550 calculus, we can write the difference quotient as

$$\sqrt{f_{0+ht}}(x) - \sqrt{f_0}(x) = \int_0^1 \frac{d}{dh} \sqrt{f_{0+ht}}(x) dh = \int_0^1 \frac{1}{2} \phi_{0+ht}(x) \sqrt{f_{0+ht}} \cdot t dh.$$

551 Therefore, by  $(a-b)^2 \leq 2a^2 + 2b^2$  and Cauchy-Schwarz inequality, we have the pointwise bound

$$\begin{aligned} & \left[ \frac{\sqrt{f_{t,\delta,z}}(x) - \sqrt{f_0}(x)}{t} - \frac{1}{2} \phi_{\delta,z}(x) \sqrt{f_0}(x) \right]^2 \\ & \leq 2 \left[ \int_0^1 \frac{1}{2} \phi_{ht,\delta,z}(x) \sqrt{f_{ht,\delta,z}} dh \right]^2 + 2 \frac{1}{2} \phi_{\delta,z}(x)^2 f_0(x) \\ & \leq \int_0^1 \frac{1}{2} \phi_{ht,\delta,z}(x)^2 f_{0+ht} dh + \phi_{\delta,z}(x)^2 f_0(x). \end{aligned}$$

552 By the generalized Lebesgue dominated convergence theorem [Roy10, p89, t19], in order to conclude  
553 (A.12), it is sufficient to show that  $\int_{\mathcal{X}} \int_0^1 \frac{1}{2} \phi_{ht,\delta,z}(x)^2 f_{0+ht} dh dx$  converges as  $t \rightarrow 0$ . By Fubini's  
554 theorem

$$\int_{\mathcal{X}} \int_0^1 \frac{1}{2} \phi_{ht,\delta,z}(x)^2 f_{0+ht} dh dx = \int_0^1 \int_{\mathcal{X}} \frac{1}{2} \phi_{ht,\delta,z}(x)^2 f_{0+ht} dx dh = \frac{1}{2} \int_0^1 I_{ht,\delta,z} dh.$$

555 Since the scores (A.11) are bounded, the information matrix  $I_{t,\delta,z}$  is continuous in  $t$  at 0, and the  
556 above integral converges to  $I_{0,\delta,z}$ .  $\square$



557 *Proof of Theorem 1.1.* By the pathwise differentiability of functional  $\theta$ , differentiability in quadratic  
 558 mean of the path  $t \rightarrow P_t^{\delta,z}$ , and Riesz' representation we have

$$\begin{aligned} \frac{d}{dt}\bigg|_{t=0} \theta(P_t^{\delta,z}) &= D\theta_P[\phi_{\delta,z}] \\ &= \int_{\mathcal{X}} \psi_P(x) \phi_{\delta,z}(x) dP. \end{aligned}$$

559 Assume that  $\psi_P(x) = \psi_P(x)\mathbb{1}_{\{f_0>0\}}(x)$ . Below  $P$  is fixed and we drop the subscript  $P$  for  
 560 convenience. Using the score  $\phi_{\delta,z}$  computed in Lemma A.2 and the fact that  $\psi$  has zero  $P$ -mean,  
 561 have the expression for the pathwise derivative as the convolution of the influence function with the  
 562 approximation to identity kernels:

$$\begin{aligned} \frac{d}{dt}\bigg|_{t=0} \theta(P_t^{\delta,z}) &= \int \psi_P(x) \left[ K^\delta(z-x)/f_0(x) - 1 \right] dP \\ &= \int_{\text{spt}P} \psi_P(x) K^\delta(z-x) dx - 0 \\ &= (\psi_P * K^\delta)(z). \end{aligned}$$

563 It suffices to show that for each  $\alpha > 0$  and  $M > 0$  the set

$$E_\alpha = \left\{ z \in \text{spt}P ; \limsup_{\delta \rightarrow 0} |(\psi_P * K^\delta)(z) - \psi_P(z)| > 2\alpha \right\}$$

564 has zero Lebesgue measure, because then  $E = \bigcup_{j=1}^{\infty} [E_{1/j} \cap \{|z| \leq j\}]$  has zero measure by  
 565 monotonicity, and the assertion (1.2) of the Theorem holds at all points  $z \in E^c$ . Thus, we may  
 566 assume that  $\psi_P$  has compact support and therefore belongs to  $L^1(\mathbb{R}^d)$ .

567 Because  $K^\delta$  is a bounded probability density function, with support in  $|x| \leq \delta$  by the dilation  
 568 construction (A.7), we can write

$$\begin{aligned} |(\psi_P * K_\delta)(z) - \psi_P(z)| &= \left| \int_{\mathbb{R}^d} [\psi_P(z-x) - \psi_P(z)] K_\delta(x) dx \right| \\ &\leq \int_{\mathbb{R}^d} |\psi_P(z-x) - \psi_P(z)| K_\delta(x) dx \\ &\leq \frac{c}{\delta^d} \int_{|x| \leq \delta} |\psi_P(z-x) - \psi_P(z)| dx. \end{aligned}$$

569 Fix  $\alpha > 0$  and recall that continuous functions of compact support are dense in  $L^1(\mathbb{R}^d)$  [SS09, p71], so  
 570 that for each  $\epsilon > 0$  we can choose a function  $g$  with  $\|\psi_P - g\|_{L^1(\mathbb{R}^d)} < \epsilon$ . By the triangle inequality  
 571 we can upper bound the expression above with

$$\frac{c}{\delta^d} \int_{|x| \leq \delta} |\psi_P(z-x) - g(z-x)| dx + \frac{c}{\delta^d} \int_{|x| \leq \delta} |g(z-x) - g(z)| dx + c'|g(z) - \psi_P(z)|.$$

572 By the continuity of  $g$  it follows that

$$\lim_{\delta \rightarrow 0} \frac{c}{\delta^d} \int_{|x| \leq \delta} |g(z-x) - g(z)| dx = 0, \quad \text{for all } z.$$

573 We find that

$$\limsup_{\delta \rightarrow 0} |(\psi_P * K_\delta)(z) - \psi_P(z)| \leq c'|\psi_P - g|^*(z) + c'|g(z) - \psi_P(z)|,$$

574 where the superscript  $*$  indicates the Hardy-Littlewood maximal function:

$$f^*(x) := \sup_{B \ni x} \frac{1}{\mathcal{L}^d[B]} \int_B |f(y)| dy, \quad \text{for } f \in L^1(\mathbb{R}^d), \quad x \in \mathbb{R}^d. \quad (\text{A.13})$$

If we set

$$F_\alpha = \{z \in \text{spt}P ; |\psi_P - g|^*(z) > \alpha\} \quad \text{and} \quad G_\alpha = \{z \in \text{spt}P ; |\psi_P(z) - g(z)| > \alpha\}$$

then  $E_\alpha \subset F_\alpha \cup G_\alpha$  by De Morgan's law since  $E_\alpha^c \supset F_\alpha^c \cap G_\alpha^c$ . Furthermore, by Chebyshev's inequality

$$\mathcal{L}^d[G_\alpha] \leq \frac{1}{\alpha} \|\psi_{P_0} - g\|_{L^1(\mathbb{R}^d)},$$

and by the Hardy-Littlewood maximal inequality [SS09, p101]

$$\mathcal{L}^d[F_\alpha] \leq \frac{3^d}{\alpha} \|\psi_{P_0} - g\|_{L^1(\mathbb{R}^d)}.$$

Recall that the function  $g$  was chosen such that  $\|\psi_{P_0} - g\|_{L^1(\mathbb{R}^d)} < \epsilon$ , so that

$$\mathcal{L}^d[E_\alpha] \leq c' \frac{3^d}{\alpha} \epsilon + c' \frac{1}{\alpha} \epsilon.$$

575 Since  $\epsilon > 0$  is arbitrary, we conclude that  $\mathcal{L}^d[E_\alpha] = 0$  and consequently  $P[\bigcup_{j=1}^\infty E_{1/j}] = 0$ .  $\square$

## 576 B Spectral representation

### 577 B.1 Calculation for Lemma 2.1

578 The characterization of the influence function as a constrained optimizer was discussed in Appendix  
579 A around equation (A.5).

580 We verify the equivalence of the constrained problem and the penalized problem via an explicit  
581 calculation that is simple and instructive for the calculation of the spectral representation. Define the  
582 penalized objective function with penalty loading  $\lambda_{\text{pen}} > 0$ :

$$J^\lambda(u) := D\theta_P[u] + \lambda_{\text{pen}} \|u\|_{L^2(P)}^2, \quad u \in L_0^2(P). \quad (\text{B.1})$$

583 Observe that  $J^\lambda$  is strictly convex on  $L_0^2(P)$  by the Cauchy-Schwarz inequality. By the strict  
584 convexity, the unique minimum of  $J$  is attained at the tangent vector  $u_0 \in L_0^2(P)$  where the  
585 derivative functional of  $J^\lambda$  vanishes [Lue97]:

$$DJ_{u_0}^\lambda[v] = 0 \quad \text{for all } v \in L_0^2(P). \quad (\text{B.2})$$

586 To compute the derivative of  $J^\lambda$  at some  $u$ , fix a direction  $v \in L_0^2(P)$  of perturbation and compute  
587 the difference quotient

$$\begin{aligned} J^\lambda(u + \epsilon v) - J^\lambda(u) &= \{D\theta_P[u + \epsilon v] + \lambda_{\text{pen}} \|u + \epsilon v\|_{2,P}^2\} - \{D\theta_P[u] + \lambda_{\text{pen}} \|u\|_{2,P}^2\} \\ &= \{\langle u + \epsilon v, \psi \rangle_{2,P} + \lambda_{\text{pen}} \langle u + \epsilon v, u + \epsilon v \rangle_{2,P}\} \\ &\quad - \{\langle u, \psi \rangle_{2,P} + \lambda_{\text{pen}} \langle u, u \rangle_{2,P}\} \\ &= \epsilon \{\langle v, \psi \rangle_{2,P} + 2\lambda_{\text{pen}} \langle v, u \rangle_{2,P}\} + O(\epsilon^2). \end{aligned} \quad (\text{B.3})$$

588 We find that the gradient (Riesz representer) of the derivative functional of  $J^\lambda$  at vector  $u$  is

$$\nabla J^\lambda(u) = \psi + 2\lambda_{\text{pen}} u. \quad (\text{B.4})$$

589 Using Riesz' representation, the first order condition (B.2) becomes

$$\begin{aligned} 0 &= DJ_{u_0}^\lambda[v] = \langle v, \nabla J^\lambda(u) \rangle_{2,P} \\ &= \langle v, \psi_P + 2\lambda_{\text{pen}} u_0 \rangle_{2,P} \quad \text{for all } v \in L_0^2(P). \end{aligned}$$

590 and conclude that  $u_0 = -\psi/2\lambda_{\text{pen}}$  is the minimizer of  $J^\lambda$ .

591 From this explicit solution to the penalized problem (2.1) we see that the direction of solution is  
592 always along the influence function, larger penalty loading  $\lambda_{\text{pen}}$  leads to solution with a smaller  
593  $L^2(P)$  norm, and  $\lambda_{\text{pen}}^* = 1/2$  uniquely identifies the influence function.

## 594 B.2 Calculation for Lemma 2.2

595 We define the projection of the influence function  $\psi_P$  on the ball  $B_M$  as the solution of the constrained  
 596 optimization program. Define the penalized objective function  $J^\lambda$  on  $L_0^2(P)$  with the regularization  
 597 loading  $\lambda \geq 0$ :

$$J^\lambda(u) := D\theta_P[u] + \lambda_{\text{pen}}\|u\|_{2,P}^2 + \lambda_{\text{reg}}\|u\|_{L^2(P)}^2, \quad u \in L_0^2(P). \quad (\text{B.5})$$

598 We observe that the constrained problem has the linear objective  $D\theta_P[v]$ , that the constraints are  
 599 given by quadratic functionals and that the problem satisfies Slater's condition and that the strong  
 600 convex duality holds.

601 The penalized objective  $J^\lambda$  is strictly convex and the first order optimality condition

$$DJ_{u_0}^\lambda[v] = 0, \quad \text{for all } v \in L^2 \quad (\text{B.6})$$

602 is necessary and sufficient. Compute the difference quotient:

$$\begin{aligned} & J^\lambda(u + \epsilon v) - J^\lambda(u) \\ &= \{D\theta_P[u + \epsilon v] + \lambda_{\text{pen}}\|u + \epsilon v\|_{2,P}^2 + \lambda_{\text{reg}}\|u + \epsilon v\|_H^2\} \\ &\quad - \{D\theta_P[u] + \lambda_{\text{pen}}\|u\|_{2,P}^2 + \lambda_{\text{reg}}\|u\|_H^2\} \\ &= \{\langle u + \epsilon v, \psi \rangle_{2,P} + \lambda_{\text{pen}}\langle u + \epsilon v, u + \epsilon v \rangle_{2,P} + \lambda_{\text{reg}}\langle u + \epsilon v, u + \epsilon v \rangle_H\} \\ &\quad - \{\langle u, \psi \rangle_{2,P} + \lambda_{\text{pen}}\langle u, u \rangle_{2,P} + \lambda_{\text{reg}}\langle u, u \rangle_H\} \\ &= \epsilon\{\langle v, \psi \rangle_{2,P} + 2\lambda_{\text{pen}}\langle v, u \rangle_{2,P} + 2\lambda_{\text{reg}}\langle v, u \rangle_H\} + O(\epsilon^2). \end{aligned} \quad (\text{B.7})$$

603 Take the limit as  $\epsilon \rightarrow 0$  to obtain:

$$\partial_v J^\lambda(u) = \langle v, \tilde{\theta} \rangle_{2,P} + 2\lambda_{\text{pen}}\langle v, u \rangle_{2,P} + 2\lambda_{\text{reg}}\langle v, u \rangle_H. \quad (\text{B.8})$$

604 From the first order condition, as  $\lambda_{\text{reg}} \rightarrow 0$ , the optimal solution  $u_0$  converges to that of the penalized  
 605 but unregularized objective function  $J^\lambda$ .

## 606 B.3 Proof of Theorem 2.3

607 First we check that the relationship (2.4) between the inner products of  $L_0^2(P)$  and  $H$  actually follows  
 608 from the assumptions about the bases. Suppose  $\{e_j\}$  and  $\{\sqrt{\sigma_j}e_j\}$  are orthonormal bases (ONB) for  
 609  $L_0^2(P)$  and  $H$  respectively and the operator  $S : L_0^2(P) \rightarrow H$  is defined by (2.3). From the definition  
 610 of the adjoint  $S^* : H \rightarrow L_0^2(P)$

$$\langle Se_j, e_i \rangle_H = \langle e_j, S^*e_i \rangle_{2,P} \quad \text{all } i, j. \quad (\text{B.9})$$

611 By the ONB assumption,

$$1 = \langle e_i, e_i \rangle_{2,P} = \langle \sqrt{\sigma_i}e_i, \sqrt{\sigma_i}e_i \rangle_H \quad \text{all } i. \quad (\text{B.10})$$

612 On the other hand, applying the eigenfunction property to (B.10) and using bilinearity of the inner  
 613 product

$$\langle e_i, e_i \rangle_{2,P} = \langle \sigma_i e_i, e_i \rangle_H = \langle Se_i, e_i \rangle_H \quad \text{all } i. \quad (\text{B.11})$$

614 Similarly, we check for  $i \neq j$ ,

$$0 = \langle \frac{1}{\sqrt{\sigma_i}}Se_i, \sqrt{\sigma_j}e_j \rangle_H = \frac{\sqrt{\sigma_j}}{\sqrt{\sigma_i}}\langle e_i, S^*e_j \rangle_{2,P} \quad \text{all } i \neq j. \quad (\text{B.12})$$

615 Since  $\sigma_j/\sigma_i \neq 0$  and  $\{e_i\}$  is complete, it follows that  $e_j$  is an eigenfunction of  $S^*$ , and in the view  
 616 of (B.11), the eigenvalue is 1 so that  $S^*[e_j]$  must be equal to  $e_j$ . In other words,  $S^*$  is the inclusion  
 617 operator  $H \rightarrow L_0^2(P)$ .

618 Next, we use the adjoint relationship (B.9) and (2.4) in the expression (B.8) for the directional  
 619 derivative of the objective function  $J$ :

$$\partial_v J(u) = \langle v, \psi_P \rangle_{2,P} + 2\lambda_{\text{pen}}\langle v, u \rangle_{2,P} + 2\lambda_{\text{reg}}\langle v, u \rangle_H \quad (\text{B.13})$$

$$= \langle v, S\psi_P \rangle_H + 2\lambda_{\text{pen}}\langle v, Su \rangle_H + 2\lambda_{\text{reg}}\langle v, u \rangle_H. \quad (\text{B.14})$$

620 It follows that the  $H$  gradient (the representer in Riesz' representation for Hilbert spaces) of  $DJ_u$  is  
 621 given by:

$$\delta_H J(u) = S[\psi_P] + 2\lambda_{\text{pen}} S[u] + 2\lambda_{\text{reg}} u \quad (\text{B.15})$$

$$= \sum_j \left\{ \sigma_j \langle \psi, e_j \rangle_{2,P} + 2\sigma_j \lambda_{\text{pen}} \langle u, e_j \rangle_{2,P} + 2\lambda_{\text{reg}} \langle u, e_j \rangle_{2,P} \right\} e_j. \quad (\text{B.16})$$

622 With this expansion of the gradient  $\delta_H J(u)$ , the first order condition

$$\delta_H J(\psi_\lambda) = 0, \quad \psi_\lambda \in H \quad (\text{B.17})$$

623 of the penalized and regularized optimization program (2.2) becomes the follow system of equations:

$$0 = \sigma_j \langle \psi, e_j \rangle_{L^2(P)} + 2\sigma_j \lambda_{\text{pen}} \langle \psi_\lambda, e_j \rangle_{L^2(P)} + 2\lambda_{\text{reg}} \langle \psi_\lambda, e_j \rangle_{L^2(P)}, \quad j \geq 1. \quad (\text{B.18})$$

624 Solving for the  $L_0^2(P)$  Fourier coefficients of the optimal solution  $\psi_\lambda$ :

$$\langle \psi_\lambda, e_j \rangle_{L^2(P)} = \frac{\sigma_j}{2\sigma_j \lambda_{\text{pen}} + 2\lambda_{\text{reg}}} \langle \psi, e_j \rangle_{L^2(P)}, \quad j \geq 1. \quad (\text{B.19})$$

625 Conclude that the optimal solution of (2.2) has the following Fourier series representation

$$\psi_{P,\lambda}(x) = \sum_{j=1}^{\infty} \frac{1}{2\lambda_{\text{pen}} + 2\lambda_{\text{reg}}/\sigma_j} \left[ \langle \psi_P, e_j \rangle_{L^2(P)} \right] e_j(x). \quad (\text{B.20})$$

626 Observe that the sequence of  $L^2(P)$  coefficients is shrunk toward zero by the eigenvalue sequence  
 627  $\{\sigma_j\}$  and is in fact a valid sequence of coefficients for an element in  $H$ .

628 Finally, recall that  $\langle \psi_P, e_j \rangle_{L^2(P)} = D\theta_P[e_j]$  and that the penalty loading should be  $\lambda_{\text{pen}} = 1/2$  for  
 629 the correct scaling of the influence function from Lemma 2.1.

## 630 C Nyström method

631 Our proofs of Lemmas 3.4, 3.5, 3.6 are modifications of [RBD10, Thm7, Prop10, Thm12].

### 632 C.1 Proof of Lemma 3.4

633 Define the sequence of random operators  $\xi_i : H \rightarrow H$  given by

$$\xi_i[\phi] = \langle \phi, k_{X_i} \rangle_H k_{X_i} - T_H[\phi], \quad \phi \in H, \quad i = 1, \dots, n. \quad (\text{C.1})$$

634 We compute the norm of the continuous operator: for any orthonormal basis  $\{\phi_j\}_{j \geq 1}$  of the rkHS  $H$

$$\begin{aligned} \|T_H\|_{\text{HS}}^2 &= \sum_{j \geq 1} \|T_H \phi_j\|_H^2 \\ &= \sum_{j \geq 1} \left\| \int_{\mathcal{X}} \phi_j(x) k_x \, dP(x) \right\|_H^2 \\ &= \sum_{j \geq 1} \left\langle \int_{\mathcal{X}} \phi_j(x) k_x \, dP(x), \int_{\mathcal{X}} \phi_j(x) k_x \, dP(x) \right\rangle_H \\ &= \sum_{j \geq 1} \int_{\mathcal{X}} \int_{\mathcal{X}} \langle \phi_j(x) k_x, \phi_j(y) k_y \rangle_H \, dP(x) dP(y) \\ &= \sum_{j \geq 1} \int_{\mathcal{X}} \int_{\mathcal{X}} \phi_j(x) \phi_j(y) K(x, y) \, dP(x) dP(y) \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \left\{ \sum_{j \geq 1} \phi_j(x) \phi_j(y) \right\} K(x, y) \, dP(x) dP(y) \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \left\{ K(x, y) \right\} K(x, y) \, dP(x) dP(y) = \|K\|_{L^2(P \otimes P)}^2 \end{aligned}$$

where we exchanged the Bochner integral with the inner product by Bochner integrability, used the reproducing property of the kernel, and the standard Mercer expansion of the kernel in the orthonormal basis that converges uniformly, exchanged the sum with the double integral by Fubini's .

Compute the Hilbert-Schmidt norm of the empirical operator, noting that  $k_{X_i}$  is the eigenfunction of the rank-1 operator:

$$\|\xi_i\|_{\text{HS}} \leq \|\phi(X_i)k_{X_i}\|_{\text{HS}} + \|T_H\|_{\text{HS}} \leq |K(X_i, X_i)| + \|K\|_{L^2(P \otimes P)} \leq 2\kappa.$$

This norm is an integrable real-valued random variable and therefore  $\xi_i$  is Bochner integrable with the expectation

$$E[\xi_i] = \int_{\mathcal{X}} \langle \cdot, k_x \rangle k_x dP(x) - T_H = 0.$$

By the strong law of large numbers for a random sequences in a separable Hilbert space (the space of Hilbert-Schmidt operators on  $H$  in our case) [Bos00, Thm2.4]

$$\|T_n - T_H\|_{\text{HS}} = \left\| \frac{1}{n} \sum_{i=1}^n \xi_i \right\|_{\text{HS}} \rightarrow 0 \quad a.s.$$

Furthermore, applying the Hoeffding inequality for bounded (in norm, as verified above) random elements of a separable Hilbert space (the space of Hilbert-Schmidt operators on  $H$ ) [Pin12], obtain

$$\left\| \frac{1}{n} \sum_{i=1}^n \xi_i \right\|_{\text{HS}} \leq \frac{2\kappa\sqrt{2\tau}}{\sqrt{n}} \quad (\text{C.2})$$

with probability at least  $1 - 2e^{-\tau}$ .

## C.2 Proof of Lemma 3.5

Applying [Kat87] to the empirical operator  $B = T_n$  and the population counterpart operator  $A = T_H$  defined on the separable rKHs  $H$ , for any nonnegative convex function  $\Phi$  with  $\Phi(0) = 0$ :

$$\sum_{j \geq 1} \Phi(\hat{\sigma}_j - \sigma_j) \leq \sum_{j \geq 1} \Phi(\gamma_j)$$

where  $\{\gamma_j\}_{j \geq 1}$  is an extended by zero enumeration of the eigenvalues of the random operator

$$B - A = T_n - T_H = \frac{1}{n} \sum_{i=1}^n \xi_i$$

defined in equation (C.1).

We apply [Kat87] with the choice  $\Phi(s) = |s|^p$  for  $p \geq 1$ . In particular, with  $p = 2$ , this becomes

$$\sum_{j \geq 1} |\hat{\sigma}_j - \sigma_j|^2 \leq \sum_{j \geq 1} |\gamma_j|^2 = \|T_n - T_H\|_{\text{HS}}^2 \leq \frac{(2\kappa)^2 2\tau}{n}$$

with probability at least  $1 - 2e^{-\tau}$  from the bound (C.2).

Recalling that the sup norm is the limit of the  $p$ -norms:

$$\begin{aligned} \sup_{j \geq 1} |\hat{\sigma}_j - \sigma_j| &= \lim_{p \rightarrow \infty} \left[ \sum_{j \geq 1} (\hat{\sigma}_j - \sigma_j)^p \right]^{\frac{1}{p}} \leq \sup_{p \rightarrow \infty} \left[ \sum_{j \geq 1} |\gamma_j|^p \right]^{\frac{1}{p}} = \sup_{j \geq 1} |\gamma_j| \\ &= \|T_n - T_H\|_{\text{op}} \leq \|T_n - T_H\|_{\text{HS}} \leq \frac{2\kappa\sqrt{2\tau}}{\sqrt{n}} \end{aligned}$$

with probability at least  $1 - 2e^{-\tau}$  from the bound (C.2).

Given  $\varepsilon > 0$ , set  $\varepsilon = \frac{2\kappa\sqrt{2\tau}}{\sqrt{n}}$  and solve for  $\tau$  to obtain  $\tau = n\varepsilon^2/2(2\kappa)^2$ . Inverting the above finite sample concentration bound, find

$$P \left[ \sup_{j \geq 1} |\hat{\sigma}_j - \sigma_j| \geq \varepsilon \right] \leq 2e^{-n\varepsilon^2/2(2\kappa)^2} \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

658 For the bound on the difference of the traces, compute the trace of the empirical operator:

$$\sum_{j \geq 1} \hat{\sigma}_j = \text{tr}(T_n) = \text{tr}(\mathbf{K}_n) = \frac{1}{n} \sum_{i=1}^n K(X_i, X_i).$$

659 Compute the trace of the population analogue: for any orthonormal basis  $\{\phi_j\}_{j \geq 1}$  of the rkHS  $H$

$$\begin{aligned} \text{tr}(T_H) &= \sum_{j \geq 1} \langle T_H \phi_j, \phi_j \rangle_H \\ &= \sum_{j \geq 1} \left\langle \int_{\mathcal{X}} \phi_j(x) k_x \, dP(x), \phi_j \right\rangle_H \\ &= \sum_{j \geq 1} \int_{\mathcal{X}} \phi_j(x) \langle k_x, \phi_j \rangle_H \, dP(x) \\ &= \sum_{j \geq 1} \int_{\mathcal{X}} \phi_j(x) \phi_j(x) \, dP(x) \\ &= \int_{\mathcal{X}} \sum_{j \geq 1} \phi_j(x) \phi_j(x) \, dP(x) \\ &= \int_{\mathcal{X}} K(x, x) \, dP(x) \end{aligned}$$

660 where we interchanged the integral  $\int dP$  with the inner product by Bochner integrability, applied the  
661 reproducing property of the kernel  $k_x$ , interchanged the sum with the integral by Fubini's, and used  
662 the standard Mercer expansion of the kernel that has uniform convergence.

663 Define the centered random variables  $\zeta_i = K(X_i, X_i) - E K(X, X)$  supported on the interval  $[\kappa, \kappa]$ ,  
664 and apply the standard Hoeffding inequality [Wai19, eq2.11]:

$$\left| \sum_{j \geq 1} \hat{\sigma}_j - \sigma_j \right| = |\text{tr}(T_n) - \text{tr}(T_H)| = \left| \frac{1}{n} \sum_{i=1}^n \zeta_i \right| \leq \varepsilon$$

665 with probability at least  $1 - 2e^{-2n\varepsilon^2/(2\kappa)^2}$ .

### 666 C.3 Proof of Lemma 3.6

667 From [RBD10, prop6], for compact positive operators  $A, B$

$$\text{if } \|A - B\|_{\text{op}} \leq [\alpha_N - \alpha_{N+1}]/4, \quad \text{then } \|P_D^B - P_N^A\|_{\text{op}} \leq \frac{2}{\alpha_N - \alpha_{N+1}} \|A - B\|_{\text{op}} \quad (\text{C.3})$$

668 where  $\alpha_N$  and  $\alpha_{N+1}$  are the  $N$ th and  $(N+1)$ st distinct eigenvalues and  $P_N^A$  is the projection on the  
669 eigenspace of the top  $N$  distinct eigenvalues of  $A$ , whereas  $P_D^B$  is the projection on the eigenspace of  
670 top eigenvalues of  $B$  of the same dimension. If, in addition,  $A, B$  are Hilbert-Schmidt,

$$\text{if } \|A - B\|_{\text{HS}} \leq [\alpha_N - \alpha_{N+1}]/4, \quad \text{then } \|P_D^B - P_N^A\|_{\text{HS}} \leq \frac{2}{\alpha_N - \alpha_{N+1}} \|A - B\|_{\text{HS}}. \quad (\text{C.4})$$

671 As [RBD10, thm12] point out, a bound on the projection onto the eigenspace of a simple (multiplicity 1)  
672 eigenvalue implies a bound on the eigenfunctions: let  $\hat{\phi}, \phi$  be unit-norm and  $\langle \hat{\phi}, \phi \rangle > 0$ , then

$$\|\hat{\phi} - \phi\|_H^2 = 2(1 - \langle \hat{\phi}, \phi \rangle_H) \leq 2(1 - \langle \hat{\phi}, \phi \rangle_H^2) = \|P_{\hat{\phi}} - P_{\phi}\|_{\text{HS}}^2.$$

673 If  $2\kappa\sqrt{2\tau}/\sqrt{n} \leq [\sigma_N + \sigma_{N+1}]/4$ , then by (C.2)  $\|T_n - T_H\|_{\text{HS}} \leq [\sigma_N + \sigma_{N+1}]/4$  with probability  
674 at least  $1 - 2e^{-\tau}$ , and therefore by (C.4)

$$\|P_{\hat{H}_N} - P_N\|_{\text{HS}}^2 \leq \frac{2^2}{[\sigma_N - \sigma_{N+1}]^2} \|T_n - T_H\|_{\text{HS}}^2 \leq \frac{(2\kappa)^2 2\tau}{n} \frac{2^2}{[\sigma_N - \sigma_{N+1}]^2}.$$



675 This event occurs if  $n \geq (2\kappa)^2 2\tau(4)^2 / [\sigma_N - \sigma_{N+1}]^2$ .

676 Next, we work with the population orthonormal basis  $\{\phi_j := \sqrt{\sigma_j} e_j\}_{j=1}^\infty$  for  $H$  and extend the  
 677 population counterpart  $\{\hat{\phi}_j := \sqrt{\hat{\sigma}_j} \hat{e}_j\}_{j=1}^n$  to an orthonormal basis for  $H$ . This is possible because  
 678 there are  $n$  independent eigenvectors  $P$ -a.s. by our assumptions that  $P$  is continuous and  $\mathbf{K}_n$  is  
 679 strictly positive definite.

680 Using Parseval's identity, and then Parseval's again with the projection operators  $(I_H - P_N)$  and  $P_N$ :

$$\begin{aligned}
 \|P_{\hat{H}_N} - P_N\|_{\text{HS}}^2 &= \sum_i \|(P_{\hat{H}_N} - P_N)\phi_i\|_H^2 = \sum_i \left[ \sum_j \left| \langle (P_{\hat{H}_N} - P_N)\phi_i, \hat{\phi}_j \rangle_H \right|^2 \right] \\
 &= \sum_{i,j=1}^{r(N)} 0 + \sum_{i \geq r(N)+1} \left[ \sum_{j=1}^{r(N)} \left| \langle \phi_i, \hat{\phi}_j \rangle_H \right|^2 \right] \\
 &\quad + \sum_{i=1}^{r(N)+1} \left[ \sum_{j \geq r(N)+1} \left| \langle -\phi_i, \hat{\phi}_j \rangle_H \right|^2 \right] + \sum_{i,j \geq r(N)+1} 0 \\
 &= \sum_{j=1}^{r(N)} \left[ \sum_{i \geq r(N)+1} \left| \langle \phi_i, \hat{\phi}_j \rangle_H \right|^2 \right] + \sum_{j \geq r(N)+1} \left[ \sum_{i=1}^{r(N)+1} \left| \langle \phi_i, \hat{\phi}_j \rangle_H \right|^2 \right] \\
 &= \sum_{j=1}^{r(N)} \left[ \sum_i \left| \langle (I - P_N)[\phi_i], \hat{\phi}_j \rangle_H \right|^2 \right] + \sum_{j \geq r(N)+1} \left[ \sum_i \left| \langle P_N[\phi_i], \hat{\phi}_j \rangle_H \right|^2 \right] \\
 &= \sum_{j=1}^{r(N)} \left[ \sum_i \left| \langle \phi_i, (I - P_N)[\hat{\phi}_j] \rangle_H \right|^2 \right] + \sum_{j \geq r(N)+1} \left[ \sum_i \left| \langle \phi_i, P_N[\hat{\phi}_j] \rangle_H \right|^2 \right] \\
 &= \sum_{j=1}^{r(N)} \left\| (I - P_N)[\hat{\phi}_j] \right\|_H^2 + \sum_{j \geq r(N)+1} \left\| P_N[\hat{\phi}_j] \right\|_H^2 \\
 &\geq \sum_{j=1}^{r(N)} \left\| (I - P_N)[\hat{\phi}_j] \right\|_H^2 + \sum_{j=r(N)+1}^n \left\| P_N[\hat{\phi}_j] \right\|_H^2
 \end{aligned}$$

681 Note that the bound we obtain a bound in terms of the rkHs norm, which implies a counterpart bound  
 682 for the  $L^2(P)$  norm.

#### 683 C.4 Proof of Theorem 3.7

684 Fix  $r \geq 1$  and  $\lambda \geq 0$ , for  $j = 1, \dots, r$ , the  $\hat{\sigma}_j \xrightarrow{P} \sigma_j$  by Lemma 3.5. Assuming for simplicity that  
 685 the eigenvalues are distinct,  $\|\hat{e}_j - e_j\|_H \xrightarrow{P} 0$  by Lemma 3.6. Recall that also  $\hat{e}_j \rightarrow e_j$  uniformly on  
 686 the compact set  $\mathcal{X}$ . Assuming  $f, \hat{f}$  are continuous and  $\hat{f} \rightarrow f$   $P$ -a.s. and  $\hat{f}/f$  is bounded on  $\text{spt}(f)$ ,  
 687 assuming that  $\psi_{\hat{f}} \rightarrow \psi_f$  in  $L^1(f)$ ; then by dominated convergence

$$\begin{aligned}
 \langle \psi_{\hat{f}}, \hat{e}_j \rangle_{L^2(\hat{f})} &= \int_{\mathcal{X}} \psi_{\hat{f}} \hat{e}_j \hat{f} \, d\mathcal{L}^d = \int_{\mathcal{X}} \psi_{\hat{f}} \hat{e}_j \hat{f} / f \, dP \\
 &\xrightarrow{P} \int_{\mathcal{X}} \psi_f e_j \, dP = \langle \psi_f, e_j \rangle_{L^2(f)}.
 \end{aligned}$$

688 Conclude that  $\|\hat{\psi}_{\lambda}^r - \psi_{\lambda}^r\|_H \rightarrow 0$  in  $P$ . For  $r_n \rightarrow \infty$  slow enough, also have  $\|\hat{\psi}_{\lambda}^{r(n)} - \psi_{\lambda}^{r(n)}\|_H \rightarrow 0$   
 689 in  $P$ . Finally, by the universality of  $H$ , there exists a sequence  $\lambda_n \rightarrow 0$  slowly enough such that  
 690  $\|\hat{\psi}_{\lambda(n)}^{r(n)} - \psi\|_{L^2(P)} \rightarrow 0$  in  $P$ .

## 691 C.5 Toy Monte Carlo experiment

692 We check our theoretical results with a simple numerical experiment. Let  $\theta(P) = E_P[X]$  be the  
693 mean functional. Then  $\psi_P(x) = x - \theta(P)$ . We use the Gaussian PSD kernel from our Example  
694 3.3 and set the shape parameter  $\epsilon = 1$ . We simulate Monte Carlo data from the standard Normal  
695 distribution, corresponding to the shape parameter  $\alpha = 1/\sqrt{2}$  of our Example 3.3. This allows us to  
696 compute the oracle  $\psi_\lambda^r$  using Hermit polynomials that we numerically evaluate using the MATLAB  
697 code provided with the textbook [FM15]. We estimate the eigenvalues  $\sigma_j$  and eigenfunctions  $e_j(X_i)$   
698 the using Nyström method via MATLAB's `eig` function. We estimate the pathwise derivatives as  
699  $\frac{1}{n} \sum_{i=1}^n X_i \hat{e}_j(X_i)$ , note this does not take into account estimation of the density and evaluation of  
700 the mean functional on the estimated distribution.