Ann Inst Stat Math (2018) 70:717–744
https://doi.org/10.1007/s10463-017-0609-x

Pointwise convergence in probability of general smoothing splines

Matthew Thorpe[1] · **Adam M. Johansen**[2]

Received: 19 January 2016 / Revised: 8 December 2016 / Published online: 4 April 2017

Abstract Establishing the convergence of splines can be cast as a variational problem which is amenable to a Γ-convergence approach. We consider the case in which the regularization coefficient scales with the number of observations, n, as $\lambda_n = n^{-p}$. Using standard theorems from the Γ-convergence literature, we prove that the general spline model is consistent in that estimators converge in a sense slightly weaker than weak convergence in probability for $p \leq \frac{1}{2}$. Without further assumptions, we show this rate is sharp. This differs from rates for strong convergence using Hilbert scales where one can often choose $p > \frac{1}{2}$.

Keywords Variational methods · Γ-convergence · Pointwise convergence · General spline model · Nonparametric smoothing

1 Introduction

Given a Hilbert space, $\mathcal{H}$, with dual $\mathcal{H}^*$, the general spline problem (Kimeldorf and Wahba 1971; Wahba 1990) is to recover $\mu^\dagger \in \mathcal{H}$ from observations, $\{(L_i, y_i)\}_{i=1}^n \subseteq \mathcal{H}^* \times \mathbb{R}$, and the model

$$y_i = L_i \mu^\dagger + \epsilon_i, \tag{1}$$

where ϵ_i and L_i are independent random variables taking values in $\mathbb{R}$ and $\mathcal{H}^*$, respectively. We assume that $\mathcal{H}$ can be decomposed into $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where, for $l = 0, 1$, $(\mathcal{H}_l, \|\cdot\|_l)$ are themselves both Hilbert spaces. For example, one may apply the

✉ Matthew Thorpe
mthorpe@andrew.cmu.edu

[1] Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

theory to the special spline problem (also referred to as smoothing splines) where $\mathcal{H} = H^m([0,1])$ ($m \geq 1$) is the Sobolev space of degree m and the observation operators are of the form $L_i\mu = \mu(t_i)$ in which t_i is sampled from some distribution over $[0,1]$. Throughout this paper, we refer to (1) as the general spline model when $L_i \in \mathcal{H}^*$ and $\mathcal{H}$ is any Hilbert space, and the special spline model when L_i is the pointwise evaluation operator and $\mathcal{H} = H^m$.

Establishing convergence and the rate of convergence of estimates μ^n of $\mu^\dagger$ remains a current area of research (Bissantz et al. 2004, 2007; Claeskens et al. 2009; Hall and Opsomer 2005; Kauermann et al. 2009; Lai and Wang 2013; Lukas 2006; Wang et al. 2011). These results establish strong convergence, in the sense of convergence with respect to a norm, and related rates of the special spline problem. Convergence with respect to the norm in the original space is typically not achievable so convergence results are in weaker topologies (equivalently larger spaces). This paper fills a gap in the literature by establishing the convergence of the general spline problem in the original space in the sense that $\forall F \in \mathcal{H}^*$, $F(\mu^n)$ converges in probability to $F(\mu^\dagger)$. There exist results for pointwise convergence of the special spline problem with equally spaced ($t_i = \frac{i}{n}$) data points (Li and Ruppert 2008; Shen and Wang 2011; Xiao et al. 2012; Yoshida and Naito 2012, 2014). Our results do not assume data points are equally spaced (we do however require that they are iid) and we consider the general case where L_i are bounded and linear operators (not necessarily pointwise evaluation).

We assume that $\dim(\mathcal{H}_0) = m < \infty$ and $\dim(\mathcal{H}_1) = \infty$. This can be seen as a multi-scale decomposition of $\mathcal{H}$. The projection of a function $\mu \in \mathcal{H}$ into the subspace $\mathcal{H}_0$ is a coarse approximation of that function. Continuing with the special spline example, one can write

$$\mu(t) = \sum_{i=0}^{m-1} \frac{\nabla^i \mu(0)}{i!} t^i + \int_0^t \frac{(t-u)^{m-1}}{(m-1)!} \nabla^m \mu(u)\, \mathrm{d}u$$

for any $\mu \in H^m$. The space $\mathcal{H}_0$ is then the space of polynomials of degree at most $m-1$. Hence $\dim(\mathcal{H}_0) = m$. Imposing a penalty on the $\mathcal{H}_1$ space, we construct a sequence of estimators μ^n of $\mu^\dagger$ as the minimizers of

$$f_n(\mu) = \frac{1}{n} \sum_{i=1}^{n} |y_i - L_i\mu|^2 + \lambda_n \|\chi_1 \mu\|_1^2$$

where $\chi_i : \mathcal{H} \to \mathcal{H}_i$ ($i = 0, 1$) is the projection of $\mathcal{H}$ onto $\mathcal{H}_i$. This paper addresses the asymptotic behaviour (as $n \to \infty$) of the general spline problem and in particular how one should choose λ_n to ensure μ^n converges (in the weak sense that $\forall F \in \mathcal{H}^*$, $F(\mu^n)$ converges in probability to $F(\mu^\dagger)$) to $\mu^\dagger$. An alternative, but closely related, method is the penalized spline problem, for example Eilers and Marx (1996), where the estimate $\mu^\dagger$ is found by minimizing f_n over functions of the form $\mu = \sum_{i=1}^{\ell} a_i B_i$ where B_i are a set of B-splines and penalizing the coefficients a_i or derivatives of μ. Typically, $\ell \ll n$, so the complexity of the problem decreases.

There are two bodies of literature on the specification of λ_n. On the one hand, there are methods which define λ_n as the minimizer of some loss function, for example

average square error. This class of techniques includes cross-validation Wahba and Wold (1975), generalized cross-validation Craven and Wahba (1979) and penalized likelihood techniques (Hastie and Tibshirani 1990; Hurvich et al. 1998; Kou and Efron 2002; Mallows 1973; Sakamoto et al. 1986; Wahba 1985). These methods provide a numerical value of λ_n for a given n and a given set of data. In the case of special splines, there are many results on the asymptotic behaviour of λ_n and μ^n for these methods, see for example (Aerts et al. 2002; Cox 1983; Craven and Wahba 1979; Li 1987; Speckman and Sun 2001; Utreras 1981, 1983; Wand 1999). The alternative approach, and the one we take in this paper, is to choose a sequence such that the estimates μ^n converge to $\mu^\dagger$ in an appropriate sense at the fastest possible rate. This strategy gives a scaling regime for λ_n, but it does not in general give specific numerical values of λ_n, i.e. it provides the optimal rate of convergence but not the associated multiplicative constant.

When considering strong convergence, many results in the literature demonstrate $\mu^n \to \mu^\dagger$ in a norm via the use of Hilbert scales—see, for example, Cox (1988), Nychka and Cox (1989), Ragozin (1983), Speckman (1985), Stone (1982) and Utreras (1985). It is not typically possible to obtain strong convergence with respect to the original norm, and it is common to resort to the use of weaker norms; for example, in the special spline problem, one starts with the space H^s but looks for convergence in L^2. The alternative, which is pursued in this paper, is to consider modes of convergence related to weak convergence in the original space, $\mathcal{H}$.

Note that for special splines strong convergence in a larger space is a weaker result than weak convergence in the original space: by the Sobolev embedding theorem, weak convergence in H^s implies strong convergence in L^2; however, the converse does not hold.

In this paper, we show that the estimators of the general spline problem converge in a sense slightly weaker than convergence weakly in probability in the large data limit, $\mu^n \rightharpoonup \mu^\dagger$, for regularization λ_n that scales to zero no faster than $n^{-\frac{1}{2}}$. In this scaling regime, we say that the general spline problem is consistent. For insufficient regularization, the spline estimators may in some sense 'blow up'. In particular, for scaling outside this regime we construct (uniformly bounded) observation operators L_i such that $\mathbb{E}\left[\|\mu^n\|^2\right] \to \infty$. Hence, without further assumptions our results are sharp.

We note that these results have practical implications. If we are interested in estimating $\mu^\dagger$ at a point t, then we let $F(\mu) = \mu(t)$ where $F \in \mathcal{H}^*$. In this setting, weak convergence, or the pointwise form considered in this paper, is the natural mode of convergence to consider. If one is interested in a global approximation of $\mu^\dagger$, then convergence of $\mu^n - \mu^\dagger$ in an appropriate norm is the more relevant. The two formulations imply different scaling results for λ_n.

There are many results in the ill-posed inverse problems literature that may be applied to the strong convergence of the general spline problem, for brevity we only mention those most relevant to this work. In Wahba (1985), two different methods of estimating λ_n were compared as $n \to \infty$ using the general spline formulation. The reproducing kernel Hilbert space setting was used in Kimeldorf and Wahba (1970) which also discussed the probabilistic interpretation behind the estimator μ^n. In Cox (1988) and Nychka and Cox (1989), the authors prove the strong convergence and

optimal rates for the spline model using an approximation $\frac{1}{n}\sum_{i=1}^{n} L_i^* L_i \approx U$ where U is compact, positive definite, self-adjoint and with dense inverse. See also Carroll et al. (1991) and Mair and Ruymgaart (1996) that consider ill-posed inverse problems without noise using similar methods. In these papers, the scaling regime for λ_n is given in terms of the rate of decay of the eigenvalues of the inverse covariance (regularization) operator $\mathcal{C}^{-1}$ (where $\|\cdot\|_1 = \|\mathcal{C}^{-1}\cdot\|_{L^2}$).

There are many more recent results addressing the asymptotic properties of splines, including (Claeskens et al. 2009; Hall and Opsomer 2005; Kauermann et al. 2009; Lai and Wang 2013; Li and Ruppert 2008; Shen and Wang 2011; Wang et al. 2011; Xiao et al. 2012; Yoshida and Naito 2012, 2014). Many of these recent results concern the asymptotics of penalized splines where one fixes the number of knot points as apposed to the smoothing spline case where the number of knots is equal to the number of data points.

It is known that the special spline problem is equivalent to a white noise problem (Brown and Low 1996). Strong convergence and rates for the white noise problem have been well studied see, for example, Agapiou et al. (2013), Bissantz et al. (2007), Goldenshluger and Pereverzev (2000) and references therein.

An interesting related result, due to Silverman (1984), gives the convergence of the smoothing kernel. That is, we can write the estimator μ^n of μ given data $\{(t_i, y_i)\}_{i=1}^n$ in the form

$$\mu^n(s) = \frac{1}{n}\sum_{i=1}^{n} K_n(s, t_i) y_i$$

for a Kernel K_n (see Lemma 8). Silverman showed that $K_n(\cdot, t)$ converges to some K uniformly on $[\epsilon, 1-\epsilon]$ for every $\epsilon > 0$ and each t (the result is valid for the special spline model and penalizing the second derivative). Whilst this result gives intuition into how the kernel behaves, it does not imply the convergence of the smoothing spline. Indeed, the convergence is not valid at the end points $\{0, 1\}$ and does not account for randomness in the observations y_i. In other words, $K_n(\cdot, t) \to K(\cdot, t)$ does not imply the convergence of μ^n (or any characterization of the limit such as we give in this paper as a solution to a variational problem). Silverman's result is, however, valid for a larger range of λ than we have here. For convergence of the kernel, it is enough that $\frac{1}{\lambda} = o(n^{2-\delta})$ for any $\delta > 0$. Our results concerning the pointwise convergence of the smoothing spline hold for λ satisfying $\frac{1}{\lambda} = O(n^{\frac{1}{2}})$.

One advantage of our approach is that we gain intuition in what happens when $\lambda_n \to 0$ too quickly. Our results show a critical rate, with respect to the scaling of λ_n, at which the methodology is ill-posed below this rate and well-posed at or above this rate. The second advantage of our approach is that, by using the Γ-convergence framework, as long as we can show that minimizers are uniformly bounded the convergence follows easily (we also need to show the Γ-limit is unique, but for our problem this is not difficult). This is easier than showing, directly, that $\mu^n - \mu^\dagger$ converges to zero. We are consequently able to employ simpler assumptions than those required by more direct arguments.

The outline of this paper is as follows. In the next section, we introduce some preliminary material. This starts by defining the notation we use in the remainder of the paper. We then remind the reader of Gâteaux derivatives, the Γ-convergence framework and the spline methodology, respectively. Section 3 contains the results for the convergence of the general spline model under appropriate conditions on the scaling in the regularization using the Γ-convergence framework. We discuss the special spline model in Sect. 4.

2 Preliminary material

2.1 Notation

We use the following standard definitions for rates of convergence.

Definition 1 We define the following.

(i) For deterministic sequences a_n and r_n, where r_n are positive and real valued, we write $a_n = O(r_n)$ if $\frac{a_n}{r_n}$ is bounded. If $\frac{a_n}{r_n} \to 0$ as $n \to \infty$, we write $a_n = o(r_n)$.

(ii) For random sequences a_n and r_n, where r_n are positive and real valued, we write $a_n = O_p(r_n)$ if $\frac{a_n}{r_n}$ is bounded in probability: for all $\epsilon > 0$ there exists M_ϵ, N_ϵ such that

$$\mathbb{P}\left(\left|\frac{a_n}{r_n}\right| \geq M_\epsilon\right) \leq \epsilon \qquad \forall n \geq N_\epsilon.$$

If $\frac{a_n}{r_n} \to 0$ in probability: for all $\epsilon > 0$

$$\mathbb{P}\left(\left|\frac{a_n}{r_n}\right| \geq \epsilon\right) \to 0 \qquad \text{as } n \to \infty$$

we write $a_n = o_p(r_n)$.

Definition 2 For deterministic positive sequences a_n and b_n, we write $a_n \lesssim b_n$ to mean there exists $M < \infty$ such that $a_n \leq M b_n$ for all n.

Throughout this paper, we say that a sequence of parameter estimators is consistent if, for any value of the "parameters" (splines in our setting), they converge in the sense made precise in Theorem 9 to the true value.

We will assume ϵ_i and L_i are independent sequences of iid random variables. Our estimators μ^n are also random variables, and therefore we can reach only probabilistic conclusions about the convergence of μ^n.

We will work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ rich enough to support a countably infinite sequence of observations $(L_i, y_i)_{i \geq 1}$. All stochastic quantifiers are taken with respect to $\mathbb{P}$ unless otherwise stated. It will be convenient to introduce the natural filtration associated with the marginal sequence (L_i), and we define for $n \in \mathbb{N}$, $\mathcal{G}_n = \sigma(L_1, \ldots, L_n)$, a sequence of sub-$\sigma$-algebras of $\mathcal{F}$. We use $\mathbb{E}[\cdot|\mathcal{G}_n]$ to denote a version of the associated conditional expectation.

To emphasize the dependence on the realization $\omega \in \Omega$, and hence of the data sequence, of our functionals, we write $f_n^{(\omega)}$.

For an operator $U : \mathcal{H} \to \mathcal{H}$, we will use $\mathrm{Ran}(U)$ to denote the range of U, i.e.

$$\mathrm{Ran}(U) = \{\mu \in \mathcal{H} : \exists \nu \in \mathcal{H} \text{ s.t. } U\nu = \mu\}.$$

When U is linear, the operator norm is defined by

$$\|U\|_{\mathcal{L}(\mathcal{H},\mathcal{H})} := \sup_{\|\mu\| \leq 1} \|U\mu\|.$$

We denote the support of a probability measure ϕ on a topological space $\mathcal{I}$ endowed with its Borel σ-algebra, by $\mathrm{supp}(\phi)$, i.e.

$$\mathrm{supp}(\phi) = \inf \left\{ \mathcal{I}' : \mathcal{I}' \subset \mathcal{I}, \mathcal{I}' \text{ is closed, and } \int_{\mathcal{I} \setminus \mathcal{I}'} \phi(\mathrm{d}t) = 0 \right\}.$$

A sequence of probability measures P_n on a Polish space is said to weakly converge to a probability measure P if for all bounded and continuous functions h we have

$$P_n h \to Ph.$$

where we write $Ph = \int h(x) \; P(\mathrm{d}x)$. If P_n weakly converges to P, then we write $P_n \Rightarrow P$.

2.2 The Gâteaux derivative

Definition 3 We say that $f : \mathcal{H} \to \mathbb{R}$ is Gâteaux differentiable at $\mu \in \mathcal{H}$ in direction $\nu \in \mathcal{H}$ if the limit

$$\partial f(\mu; \nu) = \lim_{r \to 0} \frac{f(\mu + r\nu) - f(\mu)}{r}$$

exists. We may define second-order derivatives by

$$\partial^2 f(\mu; \nu, \nu') = \lim_{r \to 0} \frac{\partial f(\mu + r\nu'; \nu) - \partial f(\mu; \nu)}{r}$$

for $\mu, \nu, \nu' \in \mathcal{H}$. Similarly for higher-order derivatives, to simplify notation, when it is clear, we write

$$\partial^s f(\mu; \nu) := \partial^s f(\mu; \nu, \ldots, \nu).$$

Theorem 4 (Taylor's Theorem) *If $f : \mathcal{H} \to \mathbb{R}$ is m times continuously Gâteaux differentiable on a convex subset $K \subset \mathcal{H}$, then, for $\mu, \nu \in K$:*

$$f(\nu) = f(\mu) + \partial f(\mu; \nu - \mu) + \frac{1}{2!}\partial^2 f(\mu; \nu - \mu, \nu - \mu) + \cdots + \frac{1}{(m-1)!}\partial^{m-1} f(\mu; \nu - \mu, \ldots, \nu - \mu) + R_m$$

where

$$R_m(\mu, \nu - \mu) = \frac{1}{(m-1)!}\int_0^1 (1-t)^{m-1}\partial^m f((1-t)\mu + t\nu; \nu - \mu)\, \mathrm{d}t.$$

2.3 Γ-convergence

Variational methods, and in particular Γ-convergence, have been used by the authors previously to prove consistency of estimators which arise as solutions to a variational problem (Thorpe and Johansen 2016; Thorpe et al. 2015). We have the following definition of Γ-convergence with respect to weak convergence.

Definition 5 (Γ-*convergence* Braides 2002, Definition 1.5) Let $\mathcal{H}$ be a Banach space. A sequence $f_n : \mathcal{H} \to \mathbb{R} \cup \{\pm\infty\}$ is said to Γ-*converge* on the domain $\mathcal{H}$ to $f_\infty : \mathcal{H} \to \mathbb{R} \cup \{\pm\infty\}$ with respect to weak convergence on $\mathcal{H}$, and we write $f_\infty = \Gamma\text{-}\lim_n f_n$, if for all $\nu \in \mathcal{H}$ we have

(i) (lim inf inequality) for every sequence (ν^n) weakly converging to ν

$$f_\infty(\nu) \leq \liminf_{n\to\infty} f_n(\nu^n);$$

(ii) (recovery sequence) there exists a sequence (ν^n) weakly converging to ν such that

$$f_\infty(\nu) \geq \limsup_{n\to\infty} f_n(\nu^n).$$

When it exists, the Γ-limit is always weakly lower semi-continuous (Braides 2002, Proposition 1.31) and therefore the minimum of the Γ-limit over weakly compact sets is achieved. An important property of Γ-convergence is that it implies the convergence of almost minimizers where μ^n is a sequence of almost minimizers of f_n if there exists a sequence δ_n with $\delta_n \to 0$ and $f_n(\mu^n) \leq \inf f_n + \delta_n$. In particular, we will make use of the following well known result which can be found in Braides (2002, Theorem 1.21).

Theorem 6 (Convergence of Minimizers) *Let $f_n : \mathcal{H} \to \mathbb{R} \cup \{\pm\infty\}$ be a sequence of functionals on a Banach space $(\mathcal{H}, \|\cdot\|)$. Assume there exists a weakly compact subset $K \subset \mathcal{H}$ with*

$$\inf_{\mathcal{H}} f_n = \inf_K f_n \quad \forall n \in \mathbb{N}.$$

If $f_\infty = \Gamma\text{-}\lim_n f_n$ and f_∞ is not identically $\pm\infty$, then

$$\min_{\mathcal{H}} f_\infty = \lim_{n\to\infty} \inf_{\mathcal{H}} f_n.$$

Furthermore, if $\mu^n \in K$ are almost minimizers of f_n, then any weak limit point minimizes f_∞.

A simple consequence of the above is the following corollary which avoids recourse to subsequences.

Corollary 7 *If in addition to the assumptions of Theorem 6 the minimizer of the Γ-limit is unique, then any sequence of almost minimizers μ^n of f_n converges weakly to the minimizer of f_∞.*

2.4 The spline framework

In this subsection, we recap the spline methodology and find an explicit representation for our estimators. In particular, we construct our estimate as a minimizer of a quadratic functional. We will show the existence and uniqueness of the minimizer.

We consider the separable Hilbert space $\mathcal{H}$ with inner product and norm given by $(\cdot,\cdot)$ and $\|\cdot\|$, respectively. We assume we can write $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $(\mathcal{H}_0, (\cdot,\cdot)_0, \|\cdot\|_0)$, $(\mathcal{H}_1, (\cdot,\cdot)_1, \|\cdot\|_1)$ are Hilbert spaces with $\dim(\mathcal{H}_0) = m$ and $\dim(\mathcal{H}_1) = \infty$. We may write

$$\|\mu\| = \|\mu\|_0 + \|\mu\|_1.$$

It is convenient to extend the domain of $||\cdot||_i$ from $\mathcal{H}_i$ to $\mathcal{H}$, setting $\|\mu\|_i := \|\chi_i \mu\| = \|\chi_i \mu\|_i$ as $\mathcal{H}_0$ is orthogonal to $\mathcal{H}_1$ by assumption. For example, in the special spline case, $\mathcal{H}_0$ is the space of polynomials of degree at most $m-1$ and $\mathcal{H}_1$ will be the space of remainder terms

$$R(t) = \mu(t) - \sum_{i=0}^{m-1} \frac{\nabla^i \mu(0)}{i!} t^i.$$

The norm on $\mathcal{H}_1$ is $\|\mu\|_1 = \|\nabla^m \mu\|_{L^2}$. Now, the projection of a function $\mu \in \mathcal{H}$ to $\mathcal{H}_1$ is just the projection $\mu \mapsto R$ given by the above expression. Clearly, $\|\mu\|_1 = \|R\|_1 = \|\chi_1 \mu\|_1$. Since $\mathcal{H}_0$ is finite dimensional, we are free to choose the norm without changing the topology; however, it is convenient to choose a norm that is orthogonal to $\mathcal{H}_1$ when viewed as a function of $\mathcal{H}$. A natural choice is $\|\mu\|_0^2 = \sum_{i=0}^{m-1} |\nabla^i \mu(0)|^2$. The special spline problem is discussed more below, particularly in Sect. 4.

We wish to estimate $\mu^\dagger \in \mathcal{H}$ given observations of the form (L_i, y_i), and L_i (as well as y_i) is random. For convenience, we summarize the general spline model in the definition below. One can also see, for example, Wahba (1990) for more details on the general spline model.

The general spline model The general spline model is given by (1) where $L_i \in \mathcal{H}^*$ are random variables and ϵ_i are iid random variables from a centred distribution,

ϕ_0, with variance σ^2. The L_i are assumed to be observed without noise and to be members of a family indexed by $\mathcal{I} \subset \mathbb{R}^d$; we write L_t to mean the operator L which depends upon a parameter $t \in \mathcal{I}$. The 'randomness' of L is characterized by the distribution, ϕ_T, of a random index $t \in \mathcal{I}$. For a sample $t_i \sim \phi_T$, we write L_i as shorthand for L_{t_i}. The operator L_i is therefore interpreted as a realization of L_{t_i}. We assume that t_i, ϵ_i are independent, and for convenience we define $\phi_{L_t\mu^\dagger}$ to be the distribution ϕ_0 shifted by $-L_t\mu^\dagger$. By the Riesz Representation Theorem, there exists $\eta_i \in \mathcal{H}$ such that $L_i\mu = (\eta_i, \mu)$ for all $\mu \in \mathcal{H}$. The sequence of observed data points $(t_1, y_1), (t_2, y_2), \ldots$ is a realization of a sequence of random elements on $(\Omega, \mathcal{F}, \mathbb{P})$. To mitigate the notational burden, we suppress the ω-dependence of t_i, y_i and L_i.

For example, in the case of special splines $L_i\mu^\dagger = \mu^\dagger(t_i)$ for some t_i a random variable distributed in $[0, 1]$. Observing L_i without noise is equivalent here to observing t_i without noise. We refer to Sect. 4 for more details.

We take our sequence of estimators μ^n of $\mu^\dagger$ as minimizers, which are subsequently shown to be unique, of $f_n^{(\omega)}$ where

$$f_n^{(\omega)}(\mu) = \frac{1}{n}\sum_{i=1}^{n}(y_i - L_i\mu)^2 + \lambda_n\|\mu\|_1^2. \tag{2}$$

By completing the square, we can easily show μ^n is given implicitly by

$$G_{n,\lambda_n}\mu^n = \frac{1}{n}\sum_{i=1}^{n} y_i\eta_i$$

where

$$G_{n,\lambda} = \frac{1}{n}\sum_{i=1}^{n}\eta_i L_i + \lambda\chi_1 \tag{3}$$

and for clarity we also suppress the ω-dependence of $G_{n,\lambda}$ from the notation. It will be necessary in our proofs to bound $\|G_{n,\lambda_n}\|_{\mathcal{H}^*}$ in terms of λ_n (for almost every sequence of observations). We do this by imposing a bound on $\|L_t\|_{\mathcal{H}^*}$ or equivalently on $\|\eta_t\|$ for almost every $t \in \mathcal{I}$. See Sect. 4 for a discussion of the special spline problem and in particular how one can find η_i. In order to bound the $\mathcal{H}_0$ norm of μ^n, we need conditions on our observation operators L_t. In particular, we will use the observation operators to define a norm on $\mathcal{H}_0$. Hence, our proofs require a uniqueness assumption of L_t in $\mathcal{H}_0$ (Assumption 3 below). It is not enough that L_t are unique over $\mathcal{H}$ as this would not necessarily contain any information on the $\mathcal{H}_0$ projection of μ^n, e.g. if $L_t\mu = L_t\chi_1\mu$ for all $\mu \in \mathcal{H}$. For clarity and future reference, we now summarize the assumptions described in the previous paragraphs.

Assumptions We make the following assumptions on $f_n^{(\omega)} : \mathcal{H} \to \mathbb{R}$ defined by (2) and $\mathcal{H}$.

1. Let $(\mathcal{H}, (\cdot,\cdot), \|\cdot\|)$ be a separable Hilbert space with $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $(\mathcal{H}_0, (\cdot,\cdot)_0, \|\cdot\|_0)$ and $(\mathcal{H}_1, (\cdot,\cdot)_1, \|\cdot\|_1)$ are Hilbert spaces. Assume $\dim(\mathcal{H}) = \dim(\mathcal{H}_1) = \infty$ and $\dim(\mathcal{H}_0) = m < \infty$.

2. The distribution of $L_i := L_{t_i}$ is specified implicitly by that of $t_i \in \mathcal{I} \subset \mathbb{R}^d$, and we assume $t_i \overset{\text{iid}}{\sim} \phi_T$.
3. We assume $|\text{supp}(\phi_T)| \geq m$ and that the L_t are unique in $\mathcal{H}_0$ in the sense that if $L_t\mu = L_r\mu$ for all $\mu \in \mathcal{H}_0$, then $t = r$.
4. There exists $\alpha > 0$ such that $\|\eta_t\| = \|L_t\|_{\mathcal{H}^*} \leq \alpha$ for ϕ_T-almost every $t \in \mathcal{I}$.

For the general spline problem, we allow multivariate regression, that is $t_i \in \mathbb{R}^d$, see for example Xiao et al. (2013, Section 7) for multivariate P-splines. However, when discussing the special spline problem, we will often assume $d = 1$ since, although our convergence results still hold for $d > 1$, there are regularity issues such as that for $2m < d$ minimizers are not automatically continuous (for $2m > d$ the Sobolev space H^m on $\mathbb{R}^d$ is embedded in C^0, this is not true for $2m < d$).

The existence of a unique minimizer to (2) is established in the following lemma.

Lemma 8 *Define $f_n^{(\omega)} : \mathcal{H} \to \mathbb{R}$ by (2) and assume $\lambda_n > 0$. Under Assumptions 1–4, the operator $G_{n,\lambda_n} : \mathcal{H} \to \mathcal{H}$ defined by (3) has a well-defined inverse G_{n,λ_n}^{-1} on span$\{\eta_1, \ldots, \eta_n\}$ for almost every $\omega \in \Omega$. In particular, there almost surely exists $N < \infty$ such that for all $n \geq N$ there exists a unique minimizer $\mu^n \in \mathcal{H}$ to $f_n^{(\omega)}$ which is given by*

$$\mu^n = \frac{1}{n}\sum_{i=1}^{n} y_i G_{n,\lambda_n}^{-1}\eta_i. \tag{4}$$

Proof We claim that any minimizer of $f_n^{(\omega)}$ lies in the set $\mathcal{H}_0 \oplus \text{span}\{\chi_1\eta_1, \ldots, \chi_1\eta_n\} =: \mathcal{H}_n'$. If this is so, and it can be shown that G_{n,λ_n}^{-1} is well defined on $\mathcal{H}_n'$, then we can conclude the minimizer must be of the form (4).

We define $\Omega' \subset \Omega$ by

$$\Omega' := \big\{\omega \in \Omega : \text{the number of unique } t_j \text{ in } \{t_i\}_{i=1}^{\infty} \text{ is greater than } m \text{ and } \|L_i\|_{\mathcal{H}^*} \leq \alpha\ \forall i\big\}.$$

By Assumptions 3 and 4, $\mathbb{P}(\Omega') = 1$. Let $\omega \in \Omega'$ and then there exists N such that for all $n \geq N$ we have that $\{L_i\}_{i=1}^N$ contains m distinct elements. Therefore, $\|\mu\|_{\mathcal{H}_n'}^2 := \frac{1}{n}\sum_{i=1}^n (L_i\mu)^2 + \lambda_n\|\mu\|_1^2$ defines a norm on $\mathcal{H}_n'$ for any $n \geq N$ and, as $\mathcal{H}_n'$ is finite dimensional, we arrive at the same topology whichever norm we choose.

We first show that any minimizer of $f_n^{(\omega)}$ lies in $\mathcal{H}_n'$. Let $\mu = \sum_{j=1}^m a_j\phi_j + \sum_{j=1}^n b_j\chi_1\eta_j + \rho$ where ϕ_j are a basis for $\mathcal{H}_0$ and $\rho \perp \mathcal{H}_n'$. Then, since $L_i\rho = (\eta_i, \rho) = 0$ we have:

$$f_n^{(\omega)}(\mu) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - L_i\chi_{\mathcal{H}_n'}\mu\right)^2 + \lambda_n\left\|\sum_{j=1}^{n} b_j\chi_1\eta_j\right\|_1^2 + \lambda_n\|\rho\|_1^2$$

where $\chi_{\mathcal{H}_n'}$ denotes the projection onto $\mathcal{H}_n'$. Trivially any minimizer of $f_n^{(\omega)}$ must have $\|\rho\|_1 = 0$ and since $\rho \in \mathcal{H}_1$ this implies $\rho = 0$. Hence, minimizers of $f_n^{(\omega)}$ lie in $\mathcal{H}_n'$.

We now show that G_{n,λ_n} has a well-defined inverse on $\mathcal{H}_n'$; that is we want to show that for any $r \in \mathcal{H}_n'$ there exists $\mu^n \in \mathcal{H}_n'$ such that $G_{n,\lambda_n}\mu^n = r$. The weak formulation of $G_{n,\lambda_n}\mu^n = r$ is given by

$$B(\mu^n, \nu) = (r, \nu) \qquad \forall \nu \in \mathcal{H}_n'$$

where

$$B(\mu, \nu) = \frac{1}{n}\sum_{i=1}^{n}(L_i\mu)(L_i\nu) + \lambda_n(\chi_1\mu, \chi_1\nu).$$

Now we apply the Lax-Milgram lemma to imply there exists a unique weak solution. Clearly $B : \mathcal{H}_n' \times \mathcal{H}_n' \to \mathbb{R}$ is a bilinear form. We will show it is also bounded and coercive. As $\omega \in \Omega'$, $\|L_i\|_{\mathcal{H}^*} \leq \alpha$ and for $\mu, \nu \in \mathcal{H}_n'$ we have

$$\begin{aligned} |B(\mu, \nu)| &\leq \frac{1}{n}\sum_{i=1}^{n}|L_i\mu L_i\nu| + \lambda_n\|\mu\|_1\|\nu\|_1 \\ &\leq \alpha^2\|\mu\|\|\nu\| + \lambda_n\|\mu\|_1\|\nu\|_1 \\ &\leq \left(\alpha^2 + \lambda_n\right)\|\mu\|\|\nu\|. \end{aligned}$$

Hence, B is bounded. Similarly, for some constant c independent of μ,

$$B(\mu, \mu) = \frac{1}{n}\sum_{i=1}^{n}(L_i\mu)^2 + \lambda_n\|\mu\|_1^2 = \|\mu\|_{\mathcal{H}_n'}^2 \geq c\|\mu\|^2$$

where the inequality follows by the equivalence of norms on finite dimensional spaces. Hence, B is coercive and by the Lax-Milgram Lemma there exists a unique weak solution. We have shown that for any $r \in \mathcal{H}_n'$ there exists $\mu_n \in \mathcal{H}_n'$ such that $B(\mu^n, \nu) = (r, \nu)$ for all $\nu \in \mathcal{H}_n'$.

A strong solution follows from the equivalence of the strong and weak topology on finite dimensional spaces or alternatively from the following short calculation. We have

$$(r, \nu) = \left(\frac{1}{n}\sum_{i=1}^{n}(L_i\mu^n)\eta_i, \nu\right) + \left(\lambda_n\chi_1\mu^n, \nu\right) \qquad \forall \nu \in \mathcal{H}_n'.$$

Hence

$$\left(r - \frac{1}{n}\sum_{i=1}^{n}(L_i\mu^n)\eta_i - \lambda_n\chi_1\mu^n, \nu\right) = 0 \qquad \forall \nu \in \mathcal{H}_n'.$$

So choosing $\nu = r - \frac{1}{n}\sum_{i=1}^{n}(L_i\mu^n)\eta_i - \lambda_n\chi_1\mu^n$ implies $\|r - \frac{1}{n}\sum_{i=1}^{n}(L_i\mu^n)\eta_i - \lambda_n\chi_1\mu^n\|^2 = 0$ and therefore

$$r = \frac{1}{n}\sum_{i=1}^{n}(L_i\mu^n)\eta_i - \lambda_n\chi_1\mu^n = G_{n,\lambda_n}\mu^n.$$

As this is true for all $r \in \mathcal{H}_n'$ we can infer the existence of an inverse operator $G_{n,\lambda_n}^{-1} : \mathcal{H}_n' \to \mathcal{H}_n'$ such that $G_{n,\lambda_n}^{-1} r = \mu^n$. One can verify that G_{n,λ_n}^{-1} is linear. As $\omega \in \Omega'$ was arbitrary, the result holds almost surely. □

3 Consistency

We demonstrate consistency by applying the Γ-convergence framework. This requires us to find the Γ-limit, to show that the Γ-limit has a unique minimizer and that the minimizers of $f_n^{(\omega)}$ are uniformly bounded. The next three subsections demonstrate that each of these requirements is satisfied under the stated assumptions and allow the application of Corollary 7 to conclude the consistency of the spline model, as summarized in Theorem 9. We start by stating the remainder of the conditions employed.

Assumptions 5. We have $\lambda_n = n^{-p}$ with $0 < p \leq \frac{1}{2}$.

6. For $\nu \in \mathcal{H}$ the following relation holds:

$$\int_{\mathcal{I}} (L_t\nu)^2 \phi_T(\mathrm{d}t) = 0 \Leftrightarrow \nu = 0.$$

7. For each $\mu \in \mathcal{H}$, each $L_t\mu$ is continuous in t, i.e $\|L_s - L_t\|_{\mathcal{H}^*} \to 0$ as $s \to t$.

Assumption 5 gives the admissible scaling regime in λ_n. Clearly if $p \leq 0$, then $\lambda_n \not\to 0$ and hence we expect the limit, if it even exists, to be biased towards solutions more regular than $\mu^\dagger$. We are required to show that the minimizers are bounded in probability. To do so, we show they are bounded in expectation. We will show in Theorem 11 that for $p > \frac{1}{2}$ we cannot bound minimizers in expectation; hence, it is not possible to extend our proofs for $p \notin (0, \frac{1}{2}]$. Theorem 9 holds as it does and not in expectation because the Γ-convergence framework requires μ^n to be a minimizer and as such we cannot make conclusions about the "average minimizer" since $\mathbb{E}[\mu^n|\mathcal{G}_n]$ is not a minimizer.

We will show that the second derivative of f_∞ in the direction ν is given by $\int_{\mathcal{I}}(L_t\nu)^2\ \phi_T(\mathrm{d}t)$. Assumption 6 is used to establish that f_∞ is strictly convex, and hence the minimizer is unique.

It will be necessary to show that

$$\frac{1}{n}\sum_{i=1}^{n}|L_i\mu| \to \int_{\mathcal{I}}|L_t\mu|\ \phi_T(\mathrm{d}t) \tag{5}$$

for all $\mu \in \mathcal{H}$ with probability one. We impose Assumption 7 (together with Assumption 4) to imply that $L_t\mu$ is continuous and bounded in t for all $\mu \in \mathcal{H}$ and therefore by the weak convergence of the empirical measure we infer that (5) holds for all $\mu \in \mathcal{H}$ and for almost every sequence $\{L_i\}_{i=1}^{\infty}$. In particular, we can define a set $\Omega' \subset \Omega$ independent of μ, on which (5) holds, such that $\mathbb{P}(\Omega') = 1$.

Theorem 9 *Define* $f_n^{(\omega)} : \mathcal{H} \to \mathbb{R}$ *by* (2)*. Under Assumptions 1–7, the minimizer* μ^n *of* $f_n^{(\omega)}$ *converges in the following sense: for all* $\epsilon, \delta > 0$ *and* $F \in \mathcal{H}^*$ *there exists* $N = N(\epsilon, \delta, F) \in \mathbb{N}$ *such that*

$$\mathbb{P}\left(\left|F(\mu^n) - F(\mu^\dagger)\right| \geq \epsilon\right) \leq \delta \quad \text{for } n \geq N.$$

Remark 10 We view the mode of convergence in the above theorem as a natural generalization of convergence in probability; it is weaker than convergence *weakly in probability*, which would require that the convergence of $\mu^n \to \mu^\dagger$ were uniform over $F \in \mathcal{H}^*$ and not pointwise as established in the theorem.

The following theorem shows that if $p > \frac{1}{2}$, then without imposing further assumptions it is always possible to construct observation functionals $\{L_t\}_{t\in\mathcal{I}}$ such that $\mathbb{E}\left[\|\mu^n\|^2\right] \to \infty$.

Theorem 11 *Define* $f_n^{(\omega)} : \mathcal{H} \to \mathbb{R}$ *by* (2)*, let* μ^n *be the minimizer of* $f_n^{(\omega)}$ *and take any* $\alpha > 0$ *and* $p > \frac{1}{2}$*. Take Assumptions 1–2, and assume that* $\lambda = n^{-p}$*. Then, there exists a distribution* ϕ_T *on* $\mathcal{I}$ *such that* $\|L_t\|_{\mathcal{H}^*} = \|\eta_t\| \leq \alpha$ *for almost every* $\omega \in \Omega$ *(i.e. Assumption 4 holds) and* $\mathbb{E}[\|\mu^n\|^2] \to \infty$*.*

In the special spline model, when $\lambda \to 0$, too quickly the functions μ^n begin to interpolate the data points $\{(t_i, y_i)\}_{i=1}^n$, and hence the derivative of μ^n will not stay bounded. Furthermore, when considering weak convergence, one is restricting to finite dimensional projections. It is therefore not surprising that $n^{-\frac{1}{2}}$ is the best we can do. For $p > \frac{1}{2}$ and a sequence of real-valued iid random variables X_i of finite variance (which are not identically zero), we have $n^{2p}\mathbb{E}(\frac{1}{n}\sum_{i=1}^n X_i)^2 \to \infty$. In light of this, elementary observation Theorem 11 is not surprising. The proof is given in Sect. 3.4.

3.1 The Γ-limit

We claim the Γ-limit of $f_n^{(\omega)}$, for almost every $\omega \in \Omega$, is given by

$$f_\infty(\mu) = \int_{\mathcal{I}} \int_{-\infty}^{\infty} |y - L_t\mu|^2 \, \phi_{L_t\mu^\dagger}(\mathrm{d}y) \, \phi_T(\mathrm{d}t). \tag{6}$$

Theorem 12 *Define* $f_n^{(\omega)}$, $f_\infty : \mathcal{H} \to \mathbb{R}$ *by* (2) *and* (6)*, respectively. Under Assumptions* 1–2, 5 and 7*,*

$$f_\infty = \Gamma\text{-}\lim_n f_n^{(\omega)}$$

for almost every $\omega \in \Omega$*.*

Proof We are required to show the two inequalities in Definition 5 hold with probability 1. In order to do this, we consider a subset of Ω of full measure, Ω', and show that both statements hold for every data sequence obtained from that set.

Define $g_\mu(t, y) = (y - L_t\mu)^2$. For clarity let $P(\mathrm{d}(t, y)) = \phi_T(\mathrm{d}t)\phi_{L_t\mu^\dagger}(\mathrm{d}y)$ and P_n be the empirical measure associated with the observations, i.e. for any measurable $h : \mathcal{I} \times \mathbb{R} \to \mathbb{R}$ we define $P_n h = \frac{1}{n}\sum_{i=1}^n h(t_i, y_i)$. Further, let $P_n^{(\omega)}$ denote the measure arising from the particular realization ω. Defining:

$$\Omega' = \left\{\omega : P_n^{(\omega)} \Rightarrow P\right\} \cap \left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\omega) \to \sigma^2 \text{ and } \frac{1}{n}\sum_{i=1}^n \epsilon_i(\omega) \to 0\right\},$$

then $\mathbb{P}(\Omega') = 1$ by the almost sure weak convergence of the empirical measure Dudley (2002, Theorem 11.4.1) and the strong law of large numbers. Let $\omega \in \Omega'$.

We start with the lim inf inequality. Pick $\nu \in \mathcal{H}$ and let $\nu^n \rightharpoonup \nu$. By Theorem 1.1 in Feinberg et al. (2014) we have

$$\begin{aligned}
&\int_{\mathcal{I}}\int_{-\infty}^{\infty} \liminf_{n\to\infty,(t',y')\to(t,y)} g_{\nu^n}(t', y')\ P(\mathrm{d}(t, y)) \\
&\quad\le \liminf_{n\to\infty} \int_{\mathcal{I}}\int_{-\infty}^{\infty} g_{\nu^n}(t, y)\ P_n^{(\omega)}(\mathrm{d}(t, y)) \\
&\quad= \liminf_{n\to\infty} f_n^{(\omega)}(\nu^n).
\end{aligned}$$

Now, we show

$$\liminf_{n\to\infty,(t',y')\to(t,y)} g_{\nu^n}(t', y') \ge g_\nu(t, y) \tag{7}$$

which proves the lim inf inequality. Let $(t_m, y_m) \to (t, y)$ and then

$$\begin{aligned}
(g_{\nu^n}(t_m, y_m))^{\frac{1}{2}} &= |y_m - L_{t_m}\nu^n| \\
&\ge |L_{t_m}\nu^n - y| - |y_m - y| \\
&\ge |y - L_t\nu^n| - |L_{t_m}\nu^n - L_t\nu^n| - |y_m - y| \\
&\ge |y - L_t\nu^n| - \|L_{t_m} - L_t\|_{\mathcal{H}^*}\|\nu^n\| - |y_m - y|.
\end{aligned}$$

A consequence of the uniform boundedness principle is that any weakly convergent sequence is bounded, and hence there exists some $C > 0$ such that $\|\nu^n\| \le C$. It follows from the above, and Assumption 7, that

$$\liminf_{n\to\infty,m\to\infty} (g_{\nu^n}(t_m, y_m))^{\frac{1}{2}} \ge |y - L_t\nu| = (g_\nu(t, y))^{\frac{1}{2}}.$$

As our choice of sequence (t_m, y_m) was arbitrary, we can conclude that (7) holds.

For the recovery sequence, we choose $\nu \in \mathcal{H}$ and let $\nu^n = \nu$. We are required to show

$$Pg_\nu \ge \limsup_{n\to\infty}\left(P_n^{(\omega)}g_\nu + \lambda_n\|\mu\|_1^2\right) = \limsup_{n\to\infty} P_n^{(\omega)}g_\nu.$$

Since we can write

$$g_\nu(t_i, y_i) = (L_i\mu^\dagger)^2 + \epsilon_i^2 + (L_i\nu)^2 + 2\epsilon_i L_i\mu^\dagger - 2L_i\mu^\dagger L_i\nu - 2\epsilon_i L_i\nu$$

and each term is either a continuous and bounded functional, or its convergence is addressed directly by the construction of Ω', we have $P_n^{(\omega)} g_\nu \to P g_\nu$ as required. As $\omega \in \Omega'$ was arbitrary, the result holds almost surely. □

Remark 13 Note that in the above theorem we did not need a lower bound on the decay of λ_n (only that $\lambda_n \geq 0$). We only used that $\lambda_n = o(1)$.

3.2 Uniqueness of the Γ-limit

To show the Γ-limit has a unique minimizer, we show it is strictly convex. The following lemma gives the second Gâteaux derivative of f_∞. After which we conclude in Corollary 15 that the Γ-limit is unique.

Lemma 14 *Under Assumptions 1–2 define $f_\infty : \mathcal{H} \to \mathbb{R}$ by (6). Then, the first and second Gâteaux derivatives of f_∞ are given by*

$$\partial f_\infty(\mu; \nu) = 2 \int_{\mathcal{I}} \int_{-\infty}^{\infty} (L_t \mu - y) L_t(\nu) \phi_{L_t \mu^\dagger}(\mathrm{d}y) \phi_T(\mathrm{d}t)$$
$$\partial^2 f_\infty(\mu; \nu, \zeta) = 2 \int_{\mathcal{I}} (L_t \nu)(L_t \zeta) \phi_T(\mathrm{d}t).$$

Proof We first compute the first Gâteaux derivative. We have

$$\begin{aligned}
\partial f_\infty(\mu; \nu) &= \lim_{r \to 0} \int_{\mathcal{I}} \int_{-\infty}^{\infty} \frac{(y - L_t(\mu + r\nu))^2 - (y - L_t \mu)^2}{r} \phi_{L_t \mu^\dagger}(\mathrm{d}y) \phi_T(\mathrm{d}t) \\
&= 2 \int_{\mathcal{I}} \int_{-\infty}^{\infty} (L_t \mu - y) L_t(\nu) \phi_{L_t \mu^\dagger}(\mathrm{d}y) \phi_T(\mathrm{d}t) \\
&\quad + \lim_{r \to 0} r \int_{\mathcal{I}} \int_{-\infty}^{\infty} (L_t \nu)^2 \phi_{L_t \mu^\dagger}(\mathrm{d}y) \phi_T(\mathrm{d}t) \\
&= 2 \int_{\mathcal{I}} \int_{-\infty}^{\infty} (L_t \mu - y) L_t(\nu) \phi_{L_t \mu^\dagger}(\mathrm{d}y) \phi_T(\mathrm{d}t) \quad \text{recalling that } L_t \text{ is linear.}
\end{aligned}$$

The second Gâteaux derivative follows similarly.

$$\begin{aligned}
\partial^2 f_\infty(\mu; \nu, \zeta) &= \lim_{r \to 0} 2 \int_{\mathcal{I}} \int_{-\infty}^{\infty} \frac{(L_t(\mu + r\zeta) - y) L_t \nu - (L_t \mu - y) L_t \nu}{r} \\
&\quad \times \phi_{L_t \mu^\dagger}(\mathrm{d}y) \phi_T(\mathrm{d}t) \\
&= 2 \int_{\mathcal{I}} \int_{-\infty}^{\infty} (L_t \nu)(L_t \zeta) \phi_{L_t \mu^\dagger}(\mathrm{d}y) \phi_T(\mathrm{d}t) \\
&= 2 \int_{\mathcal{I}} (L_t \nu)(L_t \zeta) \phi_T(\mathrm{d}t).
\end{aligned}$$

□

Corollary 15 *Under Assumptions* 1–2 and 6, *define $f_\infty : \mathcal{H} \to \mathbb{R}$ by (6). Then, f_∞ has a unique minimizer which is achieved for $\mu = \mu^\dagger$.*

Proof It is easy to check that $\partial f_\infty(\mu^\dagger; \nu) = 0$ for all $\nu \in \mathcal{H}$. By Lemma 14 and Assumption 6, the second Gâteaux derivative satisfies $\partial^2 f_\infty(\mu; \nu) > 0$ for all $\nu \neq 0$. Then, by Taylor's Theorem (and noting that f_∞ is quadratic), for $\mu \neq \mu^\dagger$,

$$f_\infty(\mu) = f_\infty(\mu^\dagger) + \frac{1}{2}\partial^2 f_\infty(\mu^\dagger; \mu - \mu^\dagger) > f_\infty(\mu^\dagger)$$

as required. □

3.3 Bound on minimizers

In this subsection, we show that $\|\mu^n\| = O_p(1)$. The bound in $\mathcal{H}_0$ can be obtained using fewer assumptions (than the bound in $\mathcal{H}$), which is natural considering $\mathcal{H}_0$ is finite dimensional. We may choose the norm on $\mathcal{H}_0$ without changing the topology (all norms are equivalent on finite dimensional spaces). We will use

$$\|\mu\|_0 = \int_{\mathcal{I}} |L_t \mu| \phi_T(\mathrm{d}t).$$

Loosely speaking, we can then write $\|\mu^n\|_0 \lesssim f_n^{(\omega)}(\mu^n)$. The bound in $\mathcal{H}_0$ then follows if $\min f_n^{(\omega)}$ is bounded. We make this argument rigorous in Lemma 16. After this result, we concentrate on bounding μ^n in $\mathcal{H}$.

Lemma 16 *Define $f_n^{(\omega)} : \mathcal{H} \to \mathbb{R}$ by (2). Under Assumptions 1–5 and 7, the minimizers μ^n of $f_n^{(\omega)}$ are, with probability one, eventually bounded in $\mathcal{H}_0$, i.e. for almost every $\omega \in \Omega$ there exist constants $C, N > 0$ such that $\|\mu^n\|_0 \leq C$ for all $n \geq N$.*

Proof We define P and $P_n^{(\omega)}$ as in the proof of Theorem 12, let

$$\Omega' = \left\{\omega \in \Omega : P_n^{(\omega)} \Rightarrow P\right\} \cap \left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\omega) \to \sigma^2 \text{ and } \frac{1}{n}\sum_{i=1}^n |\epsilon_i(\omega)| \to P|\epsilon_1|\right\}$$

and μ^n be a minimizer of $f_n^{(\omega)}$. Assume $\omega \in \Omega'$. As

$$f_n^{(\omega)}(\mu^n) \leq f_n^{(\omega)}(\mu^\dagger) \leq \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 + \lambda_1 \|\mu^\dagger\|_1^2 \to \sigma^2 + \lambda_1 \|\mu^\dagger\|_1^2,$$

there exists N such that $f_n^{(\omega)}(\mu^n) \leq \sigma^2 + \lambda_1 \|\mu^\dagger\|_1^2 + 1$ for $n \geq N$.

Note that for any $a, b \in \mathbb{R}$ we have

$$|a-b|^2 \geq \begin{cases} |a-b| & \text{if } |a-b| \geq 1 \\ |a-b| - 1 & \text{otherwise.} \end{cases}$$

In either case $|a-b|^2 \geq |a-b|-1 \geq |a|-|b|-1$. Now

$$\begin{aligned} f_n^{(\omega)}(\mu) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - L_i\mu)^2 + \lambda_n\|\mu\|_1^2 \\ &\geq \frac{1}{n}\sum_{i=1}^{n}(|L_i\mu| - |y_i| - 1) \\ &= \frac{1}{n}\sum_{i=1}^{n}|L_i\mu| - \frac{1}{n}\sum_{i=1}^{n}|y_i| - 1 \\ &\geq \frac{1}{n}\sum_{i=1}^{n}|L_i\mu| - \frac{1}{n}\sum_{i=1}^{n}|L_i\mu^\dagger| - \frac{1}{n}\sum_{i=1}^{n}|\epsilon_i| - 1 \\ &\to \int_{\mathcal{I}}|L_t\mu|\phi_T(\mathrm{d}t) - c \end{aligned}$$

where the convergence follows since $|L_t\mu|$ is a continuous and bounded functional in t and c is given by

$$\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}|L_i\mu^\dagger| + \frac{1}{n}\sum_{i=1}^{n}|\epsilon_i| + 1\right) \leq \int_{\mathcal{I}}|L_t\mu^\dagger|\phi_T(\mathrm{d}t) + \sigma + 1 =: c.$$

We now show that $\int_{\mathcal{I}}|L_t\mu|\phi_T(\mathrm{d}t)$ is a norm on $\mathcal{H}_0$ and hence that the above constant, c, is finite. This will also show that $\|\mu\|_0 \leq f_n^{(\omega)}(\mu) + c$ for $n \geq N$, which completes the proof.

The triangle inequality, absolute homogeneity and that $\int_{\mathcal{I}}|L_t\mu|\phi_T(\mathrm{d}t) \geq 0$ are trivial to establish. By Assumption 3, we have at least m disjoint subsets of positive measure (with respect to ϕ_T) on $\mathcal{I}$. If $\int_{\mathcal{I}}|L_t\mu|\phi_T(\mathrm{d}t) = 0$ then it follows that on each of these subsets $L_t\mu = 0$. As $\mathcal{H}_0$ is m-dimensional this determines μ, and hence $\mu = 0$.

As $\omega \in \Omega'$ was arbitrary and $\mathbb{P}(\Omega') = 1$, the result holds almost surely. □

Remark 17 In the above lemma, we did not need the lower bound on λ_n (only that $\lambda_n \geq 0$). The result holds for all $\lambda_n = O(1)$.

Continuing with the bound in $\mathcal{H}$, we write

$$\mu^n = \frac{1}{n}\sum_{i=1}^{n}L_i\mu^\dagger G_{n,\lambda_n}^{-1}\eta_i + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i G_{n,\lambda_n}^{-1}\eta_i = G_{n,\lambda_n}^{-1}U_n\mu^\dagger + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i G_{n,\lambda_n}^{-1}\eta_i \quad (8)$$

where

$$U_n = \frac{1}{n}\sum_{i=1}^{n}\eta_i L_i. \quad (9)$$

We bound $\|G_{n,\lambda_n}^{-1}U_n\mu^\dagger\|$ in Lemma 19 and $\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i G_{n,\lambda_n}^{-1}\eta_i\|$ in Lemma 20.

In the proof of Lemma 19, we show that $G_{n,\lambda_n}^{-1} : \mathrm{Ran}(U_n) \to \mathrm{Ran}(U_n)$. Lemma 18 gives the conditions necessary to infer the existence of a orthonormal basis of eigenfunctions $\{\psi_j^{(n)}\}_{j=1}^{\infty}$ of $\mathrm{Ran}(U_n)$. Hence, we can write

$$\|G_{n,\lambda_n}^{-1} U_n \mu\|^2 = \sum_{j=1}^{\infty} (G_{n,\lambda_n}^{-1} U_n \mu, \psi_j^{(n)})^2.$$

From here, we exploit the fact that $\psi_j^{(n)}$ are eigenfunctions. We leave the details until the proof of Lemma 19.

Lemma 20 is a consequence of being able to bound $\|G_{n,\lambda_n}^{-1}\|_{\mathcal{L}(\mathcal{H},\mathcal{H})}$ in terms of λ_n. One is then left to show $\left(\frac{1}{n}\sum_{i=1}^{n} \epsilon_i\right)^2 = O(\frac{1}{n})$. We start by showing that U_n is compact, bounded, self-adjoint and positive semi-definite.

Lemma 18 *Define U_n by (9). Under Assumptions* 1 and 4, *U_n is almost surely a bounded, self-adjoint, positive semi-definite and compact operator on $\mathcal{H}$.*

Proof In this proof, we consider $\omega \in \Omega'$ where $\Omega' = \{\omega : \|\eta_i(\omega)\| \leq \alpha \text{ for all } i\}$, noting that $\mathbb{P}(\Omega') = 1$ by Assumption 4.

Boundedness of U_n follows easily as

$$\|U_n \mu\| \leq \frac{1}{n} \sum_{i=1}^{n} \alpha^2 \|\mu\| = \alpha^2 \|\mu\|.$$

Let $(\cdot, \cdot)_{\mathbb{R}^n}$ be the inner product on $\mathbb{R}^n$ given by

$$(x, y)_{\mathbb{R}^n} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i \qquad \forall x, y \in \mathbb{R}^n.$$

Now, for $x \in \mathbb{R}$ and $v \in \mathcal{H}$ we have

$$(x, L_i v)_{\mathbb{R}^1} = x L_i v = x(\eta_i, v) = (x\eta_i, v)$$

which shows $L_i^* : \mathbb{R} \to \mathcal{H}$ is given by $L_i^* x = x\eta_i$. Now, if we define $T_n = (L_1, \ldots, L_n) : \mathcal{H} \to \mathbb{R}^n$, then for $x \in \mathbb{R}^n$, $v \in \mathcal{H}$

$$(T_n v, x)_{\mathbb{R}^n} = \frac{1}{n} \sum_{i=1}^{n} L_i v x_i = \left(\frac{1}{n} \sum_{i=1}^{n} x_i \eta_i, v\right).$$

Hence, $T_n^* x = \frac{1}{n}\sum_{i=1}^{n} x_i \eta_i$. We have shown $U_n = T_n^* T_n$ and is therefore self-adjoint.

To show U_n is positive semi-definite, we need

$$(U_n v, v) \geq 0$$

for all $v \in \mathcal{H}$. This follows easily as

$$(U_n v, v) = \frac{1}{n}\sum_{i=1}^{n}(L_i v)^2 \geq 0.$$

For compactness of U_n (for n fixed), let v^m be a sequence with $\|v^m\| \leq 1$. Since $|L_i v^m| \leq \alpha$ for every $\omega \in \Omega'$, there exists a convergent subsequence m_p such that

$$L_i v^{m_p} \to \kappa_i \quad \forall i = 1, 2, \ldots, n \quad \text{say.}$$

So $U_n v^{m_p} \to \frac{1}{n}\sum_{i=1}^{n} \eta_i \kappa_i \in \mathcal{H}$ as $m_p \to \infty$. Therefore, each U_n is compact. □

Using the basis whose existence is implied by the previous lemma, we can bound the first term on the RHS of (8).

Lemma 19 *Under Assumptions* 1–4, *define* G_{n,λ_n} *and* U_n *by* (3) *and* (9), *respectively. Then, with probability one we have*

$$\|G_{n,\lambda_n}^{-1} U_n\|_{\mathcal{L}(\mathcal{H},\mathcal{H})} \leq 1$$

for all n.

Proof First note that $\dim(\text{Ran}(U_n)) = \dim(\text{span}\{\eta_1, \ldots, \eta_n\}) \leq n$. Without loss of generality, we will assume $\dim(\text{Ran}(U_n)) = n$ (else we can assume the dimension is m_n where $m_n \leq n$ is an increasing sequence). Clearly χ_1 is a self-adjoint, bounded and compact operator on $\text{Ran}(U_n)$ as is U_n by Lemma 18. Therefore, there exists a simultaneous diagonalization of U_n and χ_1 on $\text{Ran}(U_n)$, i.e. there exists $\beta_j^{(n)}$, $\gamma_j^{(n)}$ and $\psi_j^{(n)}$ such that

$$U_n \psi_j^{(n)} = \beta_j^{(n)} \psi_j^{(n)} \quad \text{and} \quad \chi_1 \psi_j^{(n)} = \gamma_j^{(n)} \psi_j^{(n)}$$

for all $j = 1, 2, \ldots, n$. Since χ_1 is the projection operator, we must have $\gamma_j^{(n)} \in \{0, 1\}$. Furthermore, $\psi_j^{(n)}$ form an orthonormal basis of $\text{Ran}(U_n)$. Since U_n is positive semi-definite, it follows that $\beta_j^{(n)} \geq 0$. We have

$$G_{n,\lambda_n} \psi_j^{(n)} = U_n \psi_j^{(n)} + \lambda_n \chi_1 \psi_j^{(n)} = \left(\beta_j^{(n)} + \lambda_n \gamma_j^{(n)}\right) \psi_j^{(n)}.$$

So,

$$G_{n,\lambda_n}^{-1} \psi_j^{(n)} = \frac{1}{\beta_j^{(n)} + \lambda_n} \psi_j^{(n)}.$$

In particular, this shows that

$$G_{n,\lambda_n}^{-1} U_n : \mathcal{H} \to \text{Ran}(U_n).$$

Assume $\mu \in \mathcal{H}$, $\nu \in \mathrm{Ran}(U_n)$, then

$$\mu = \sum_{i=1}^{n} (\mu, \psi_i^{(n)}) \psi_i^{(n)} + \hat{\mu} \qquad \text{and} \qquad \nu = \sum_{i=1}^{n} (\nu, \psi_i^{(n)}) \psi_i^{(n)}$$

where $\hat{\mu} \in \mathrm{Ran}(U_n)^{\perp}$. Therefore,

$$(U_n \mu, \psi_j^{(n)}) = \sum_{i=1}^{n} (\mu, \psi_i^{(n)})(U_n \psi_i^{(n)}, \psi_j^{(n)}) = \beta_j^{(n)} (\mu, \psi_j^{(n)})$$
$$(G_{n,\lambda_n}^{-1} \nu, \psi_j^{(n)}) = \sum_{i=1}^{n} (\nu, \psi_i^{(n)})(G_{n,\lambda_n}^{-1} \psi_i^{(n)}, \psi_j^{(n)}) = \frac{1}{\beta_j^{(n)} + \lambda_n \gamma_j^{(n)}} (\nu, \psi_j^{(n)})$$

which implies

$$(G_{n,\lambda_n}^{-1} U_n \mu, \psi_j^{(n)}) = \frac{1}{\beta_j^{(n)} + \lambda_n \gamma_j^{(n)}} (U_n \mu, \psi_j^{(n)}) = \frac{\beta_j^{(n)}}{\beta_j^{(n)} + \lambda_n \gamma_j^{(n)}} (\mu, \psi_j^{(n)}).$$

Hence

$$\begin{aligned} \|G_{n,\lambda_n}^{-1} U_n \mu\|^2 &= \sum_{j=1}^{n} (G_{n,\lambda_n}^{-1} U_n \mu, \psi_j^{(n)})^2 \\ &= \sum_{j=1}^{n} \left(\frac{\beta_j^{(n)}}{\beta_j^{(n)} + \lambda_n \gamma_j^{(n)}} \right)^2 (\mu, \psi_j^{(n)})^2 \\ &\leq \sum_{j=1}^{n} (\mu, \psi_j^{(n)})^2 \\ &\leq \|\mu\|^2. \end{aligned}$$

This proves the lemma. □

We now focus on bounding $\|G_{n,\lambda_n}^{-1} \nu_n\|$ where $\nu_n = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \eta_i$.

Lemma 20 *Under Assumptions* 1–5, *define* G_{n,λ_n} *by* (3). *Then*

$$\mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i G_{n,\lambda_n}^{-1} \eta_i \right\|^2 \middle| \mathcal{G}_n \right] = O(1) \quad \textit{almost surely.}$$

Proof Recalling B from the proof of Lemma 8, we have

$$(G_{n,\lambda_n} \mu, \mu) = B(\mu, \mu) \geq \lambda_n \|\mu\|_1^2.$$

This implies $\|G_{n,\lambda_n}\mu\| \geq \lambda_n \|\mu\|_1$. By Lemma 8, there exists a well-defined inverse of G_{n,λ_n} at η_i, hence we let $\mu = G_{n,\lambda_n}^{-1}\eta_i$ and we have

$$\|G_{n,\lambda_n}^{-1}\eta_i\|_1 \leq \frac{1}{\lambda_n}\|\eta_i\| \leq \frac{\alpha}{\lambda_n}.$$

Almost surely. Now, define $\nu_n = \frac{1}{n}\sum_{i=1}^n \epsilon_i \eta_i$ and

$$\mathbb{E}\left[\left\|G_{n,\lambda_n}^{-1}\nu_n\right\|_1^2 \middle| \mathcal{G}_n\right] \overset{\text{a.s.}}{=} \frac{\sigma^2}{n^2}\sum_{i=1}^n \left\|G_{n,\lambda_n}^{-1}\eta_i\right\|_1^2 \leq \frac{\alpha^2\sigma^2}{n\lambda_n^2}.$$

Combined with Lemma 16 (the $\mathcal{H}_0$ bound), this proves the lemma. □

Recalling (8) and via Lemmas 19 and 20, we obtain the following asymptotic bound on minimizers in $\mathcal{H}$.

Theorem 21 *Under Assumptions* 1–5*, we have*

$$\mathbb{E}\left[\|\mu^n\|^2 | \mathcal{G}_n\right] = O(1) \quad \textit{almost surely.} \tag{10}$$

This is a stronger result than we needed; we were only required to show that $\|\mu^n\|$ is bounded in probability. Taking expectation of (10), one has

$$\mathbb{E}\|\mu^n\|^2 = O(1).$$

Hence, applying Chebyshev's inequality, we may conclude that $\|\mu^n\| = O_p(1)$.

Corollary 22 *Under Assumptions* 1–5*, we have* $\|\mu^n\| = O_p(1)$.

We conclude this section with a brief analysis of the rate of convergence. For any $F \in \mathcal{H}^*$, by the Riesz Representation Theorem, there exists $\xi \in \mathcal{H}$ such that $F(\mu) = (\mu, \xi)$ for all $\mu \in \mathcal{H}$. Hence,

$$F(\mu^n) - F(\mu^\dagger) = ((G_{n,\lambda_n}^{-1}U_n - \mathrm{Id})\mu^\dagger + G_{n,\lambda_n}^{-1}\nu^n, \xi)$$

where $\nu^n = \frac{1}{n}\sum_{i=1}^n \epsilon_i \eta_i$. Decomposing $\mathcal{H}$ into $\mathcal{H} = \overline{\mathrm{Ran}(U_n)} \oplus \mathrm{Ran}(U_n)^\perp$, one can write

$$\begin{aligned} F(\mu^n) - F(\mu^\dagger) &= \left(\left(G_{n,\lambda_n}^{-1}U_n - \chi_{\overline{\mathrm{Ran}(U_n)}}\right)\mu^\dagger, \xi\right) \\ &\quad - \left(\chi_{\mathrm{Ran}(U_n)^\perp}\mu^\dagger, \xi\right) + \left(G_{n,\lambda_n}^{-1}\nu^n, \xi\right) \\ &= \sum_{j=1}^n \frac{-\lambda_n}{\beta_j^{(n)} + \lambda_n}\left(\mu^\dagger, \psi_j^{(n)}\right)\left(\psi_j^{(n)}, \xi\right) \\ &\quad - \left(\chi_{\mathrm{Ran}(U_n)^\perp}\mu^\dagger, \xi\right) + \left(G_{n,\lambda_n}^{-1}\nu^n, \xi\right) \end{aligned} \tag{11}$$

where $\chi_{\overline{\text{Ran}(U_n)}}$ is the projection onto $\overline{\text{Ran}(U_n)}$. If we assume

$$\lim_{n\to\infty} \sum_{j=1}^{n} \frac{1}{\beta_j^{(n)}} < \infty, \tag{12}$$

then

$$\sum_{j=1}^{n} \frac{-\lambda_n}{\beta_j^{(n)} + \lambda_n} \left(\mu^\dagger, \psi_j^{(n)}\right) \left(\psi_j^{(n)}, \xi\right) \leq \|\mu^\dagger\| \|\xi\| \lambda_n \sum_{j=1}^{n} \frac{1}{\beta_j^{(n)}}.$$

And therefore the first term in (11) is of the order n^{-p}. By the proof of Lemma 20, the third term in (11) is of order $\frac{1}{\sqrt{n}\lambda_n}$. The second term is independent of λ_n. The optimal rate of convergence is therefore found by balancing the first and third terms. This will imply an optimal choice of $p = \frac{1}{4}$. We summarize in the following proposition.

Proposition 23 *Under Assumptions* 1–6, *for* $F \in \mathcal{H}^*$ *take* $\xi \in \mathcal{H}$ *such that* $F(\mu) = (\mu, \xi)$ *and assume* (12) *holds and that there exists* $q > 0$ *such that*

$$\left| \|\mu^\dagger\| - \|\chi_{\text{Ran}(U_n)}\mu^\dagger\| \right| \lesssim n^{-q}$$

where U_n *is defined by* (9) *and* $(\beta_j^{(n)}, \psi_j^{(n)})$ *are an eigenvalue–eigenfunction pair for* U_n. *Then*

$$\mathbb{E}\left[|F(\mu^n) - F(\mu^\dagger)| \, |\mathcal{G}_n \right] = O\left(n^{-p}\right) + O\left(n^{-q}\right) + O\left(\frac{1}{\lambda_n \sqrt{n}}\right) \quad \text{a.s.} \tag{13}$$

In particular, the optimal choice is $p = \frac{1}{4}$ *in which case the rate of convergence is*

$$\mathbb{E}\left[|F(\mu^n) - F(\mu^\dagger)| \, |\mathcal{G}_n \right] = O\left(n^{\max\left\{-\frac{1}{4}, -q\right\}}\right).$$

Proof The argument preceding the theorem provides the proof for the first term in (13), and the third term is a consequence of Lemma 20. The second term follows easily from

$$\left| \left(\chi_{\text{Ran}(U_n)^\perp}\mu^\dagger, \xi\right) \right| \leq \|\xi\| \|\chi_{\text{Ran}(U_n)^\perp}\mu^\dagger\| \leq \|\xi\| \left(\|\mu^\dagger\| - \|\chi_{\text{Ran}(U_n)}\mu^\dagger\| \right).$$

The optimal rate is a consequence of choosing p that minimizes $n^{-p} + n^{p-0.5}$. □

The conditions of the above theorem are difficult to theoretically verify. Even for the special spline problem, the authors know of no method to check whether assumption (12) holds and whether such a q exists. We leave further investigation into the rate of convergence for future works.

3.4 Sharpness of the scaling regime: Proof of Theorem 11

Proof of Theorem 11 Fix any $\alpha > 0$, and without loss of generality we can choose $\{\eta_t\}_{t\in\mathcal{I}}$ such that $\|\eta_t\| = \alpha$ for all $t \in \mathcal{I}$. Define $L_t \in \mathcal{H}$ by $L_t = (\eta_i, \cdot)$.

In the proof of Lemma 8, we showed

$$\left|(G_{n,\lambda_n}\mu, \nu)\right| \leq (\alpha^2 + \lambda_n)\|\mu\|\|\nu\|.$$

Letting $\nu = G_{n,\lambda_n}\mu$, for $\mu \in \text{span}\{\eta_1, \ldots, \eta_n\}$, one has

$$\|G_{n,\lambda_n}\mu\|^2 \leq (\alpha^2 + \lambda_n)\|\mu\|\|G_{n,\lambda_n}\mu\|.$$

And hence

$$\|G_{n,\lambda_n}\mu\| \leq (\alpha^2 + \lambda_n)\|\mu\|.$$

which implies

$$\|G^{-1}_{n,\lambda_n}\mu\| \geq \frac{1}{\alpha^2 + \lambda_n}\|\mu\|.$$

Now, for $\nu^n = \frac{1}{n}\sum_{i=1}^n \epsilon_i \eta_i$, we consider

$$\begin{aligned}\mathbb{E}\left[\|G^{-1}_{n,\lambda_n}\nu^n\|^2 \,\middle|\, \mathcal{G}_n\right] &\geq \frac{1}{(\alpha^2 + \lambda_n)^2}\mathbb{E}\left[\|\nu^n\|^2|\mathcal{G}_n\right] \\ &\overset{\text{a.s.}}{=} \frac{\sigma^2\alpha^2}{\lambda_n^2 n(\alpha^2 + \lambda_n)^2} \\ &\to \infty\end{aligned}$$

as $\lambda_n^2 n \to 0$. Hence, by taking expectations:

$$\mathbb{E}\left[\|G^{-1}_{n,\lambda_n}\nu^n\|^2\right] \to \infty.$$

By noting

$$\mathbb{E}\left[\|\mu^n\|^2\right] = \mathbb{E}\left[\|G^{-1}_{n,\lambda_n}U_n\mu^\dagger\|^2\right] + \mathbb{E}\left[\|G^{-1}_{n,\lambda_n}\nu^n\|^2\right]$$

we conclude the proof. □

4 Application to the special spline model

Consider the application to the special spline case, $L_i\mu = \mu(t_i)$. We let

$$\mathcal{H} = H^m := \left\{g : [0,1] \to \mathbb{R} \text{ s.t } \nabla^i g \text{ abs. cts. for } \quad i = 1, 2, \ldots, m-1 \text{ and } \nabla^m g \in L^2\right\}.$$

For $m \geq 1$, $\mathcal{H}$ is a reproducing kernel Hilbert space and therefore L_i as defined are linear and bounded operators on $\mathcal{H}$. See Bogachev (1998) and Wahba (1990) for more details on reproducing kernel Hilbert spaces. The special spline solution is the minimizer of

$$f_n(\mu) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu(t_i))^2 + \lambda_n \|\nabla^m \mu\|_{L^2}^2$$

over all $\mu \in H^m$. It can be shown that the minimizer $\mu^{(n)}$ of f_n is a piecewise polynomial of degree $2m-1$ in each interval (t_i, t_{i+1}) for $i = 0, \ldots, n$ (where we define $t_0 = 0$ and $t_{n+1} = 1$), for example see Wahba (1990, Section 1.3).

This section discusses the following points.

1. The decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_0$ is finite dimensional.
2. The function η_t corresponding to $(\eta_t, \mu) = L_t \mu = \mu(t)$.

The other assumptions needed to apply Theorem 9 are Assumption 3 and Assumption 6. Assumption 3 is

$$\mu(t) = \mu(r) \quad \text{for all polynomials } \mu \text{ of degree at most } m-1 \text{ then } t = r$$

which clearly holds. Assumption 6 becomes

$$\int_0^1 |\nu(t)|^2 \phi_T(\mathrm{d}t) = 0 \Leftrightarrow \nu = 0$$

which, for example, is true if $\phi_T(\mathrm{d}t) = \hat{\phi}_T(t)\,\mathrm{d}t$ and $\hat{\phi}_T(t) > 0$ for all $t \in [0,1]$.
1. *The decomposition* $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ For $\mu \in \mathcal{H}$ by Taylor expanding μ from 0, we can write:

$$\mu(t) = \sum_{i=0}^{m-1} \frac{\nabla^i \mu(0)}{i!} t^i + R(t)$$

where $\nabla^i R(0) = 0$ for all $i = 0, 1, \ldots, m-1$. Hence, $R \in \mathcal{H}_1$ where

$$\mathcal{H}_1 = \left\{ g \in H^m : \nabla^i g(0) = 0 \text{ for all } i = 0, 1, \ldots, m-1 \right\}.$$

A Poincaré inequality holds on this space so $\|\mu\|_1^2 = \int_0^1 |\nabla^m \mu(t)|^2\,\mathrm{d}t$ is a norm on $\mathcal{H}_1$.

We define $\mathcal{H}_0$ to be the span of the functions ζ_i defined by

$$\zeta_i(t) = \frac{t^i}{i!} \qquad \text{for } i = 0, 1, \ldots, m-1.$$

The space is equipped with the inner product

$$(\mu, \nu)_0 = \sum_{i=0}^{m-1} \nabla^i \mu(0) \nabla^i \nu(0).$$

The space $\mathcal{H}_0$ has $\dim(\mathcal{H}_0) = m$.
2. *The functions* η_t In the above, R is given by

$$R(t) = \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} \nabla^m \mu(u)\, \mathrm{d}u = \int_0^1 G(t,u) \nabla^m \mu(u)\, \mathrm{d}u$$

where $(u)_+ = \max\{0, u\}$ and

$$G(t,u) = \frac{(t-u)_+^{m-1}}{(m-1)!}$$

is the Green's function for $\nabla^m \mu = \nu$ and boundary conditions $\nabla^j \mu(0) = 0$ for all $0 \le j \le m-1$.

We claim that $\eta_t \in H^m$ satisfying $(\eta_t, \mu) = \mu(t)$ are given by

$$\eta_t(r) = \sum_{i=0}^{m-1} \zeta_i(t)\zeta_i(r) + \int_0^1 G(t,u)G(r,u)\, \mathrm{d}u =: \eta_t^0(r) + \eta_t^1(r).$$

Furthermore, $\eta_t^0 \in \mathcal{H}_0$ and $\eta_t^1 \in \mathcal{H}_1$ for all $t \in [0,1]$. The proof follows directly from calculating

$$(\eta_t, \mu) = \sum_{i=0}^{m-1} \nabla^i \eta_t(0) \nabla^i \mu(0) + \int_0^1 \nabla^m \eta_t(u) \nabla^m \mu(u)\, \mathrm{d}u$$

and noticing

$$\nabla^i \eta_t(r) = \sum_{j=1}^{m-1} \zeta_j(t) \left[\nabla^i \zeta_j(r)\right]_{r=0} = \zeta_i(t) \quad \text{for } i < m$$

$$\nabla^m \eta_t(r) = \nabla_r^m \int_0^1 G(t,u)G(r,u)\, \mathrm{d}u = G(t,r).$$

One can easily show that $\|\eta_t\| \le 1$ for all $t \in [0,1]$.

Continuity of η_t follows easily. As each polynomial is Lipschitz continuous on the interval $[0,1]$, there exists a constant C_i (depending on the order of the polynomial i) such that $|\zeta_i(t) - \zeta_i(s)| \le C_i |t-s|$. Now for the integral term, let $m \ge 2$ and $s \ge t$ then:

$$\begin{aligned}
&\left| \int_0^1 \left(G(s,u) - G(t,u)\right) G(r,u)\, \mathrm{d}u \right| \\
&\quad = \left| \int_0^1 \left(\mathbb{I}_{s>u} \frac{(s-u)^{m-1}}{(m-1)!} - \mathbb{I}_{t>u} \frac{(t-u)^{m-1}}{(m-1)!} \right) G(r,u)\, \mathrm{d}u \right| \\
&\quad \leq \int_t^s \frac{(s-u)^{m-1}}{(m-1)!} G(r,u)\, \mathrm{d}u \\
&\qquad + \frac{1}{(m-2)!} \int_0^t |s-t|\, g(r,u)\, \mathrm{d}u \\
&\quad \leq \frac{m|s-t|}{[(m-1)!]^2}.
\end{aligned}$$

The case $m = 1$ is similar. It follows that $\|L_s - L_t\|_{\mathcal{H}^*} = \|\eta_s - \eta_t\| \leq C|s-t|$ for some $C < \infty$ and hence L_t is continuous.

Acknowledgements This work was carried out whilst MT was part of MASDOC at the University of Warwick and supported by an EPSRC Industrial CASE Award Ph.D. Studentship with Selex ES Ltd.

References

Aerts, M., Claeskens, G., Wand, M. P. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference*, *103*(1–2), 455–470.

Agapiou, S., Larsson, S., Stuart, A. M. (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Processes and their Applications*, *123*(10), 3828–3860.

Bissantz, N., Hohage, T., Munk, A. (2004). Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems*, *20*(6), 1773–1789.

Bissantz, N., Hohage, T., Munk, A., Ruymgaart, F. (2007). Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis*, *45*(6), 2610–2636.

Bogachev, V. I. (1998). *Gaussian measures*. Providence: The American Mathematical Society.

Braides, A. (2002). Γ*-convergence for beginners*. Oxford: Oxford University Press.

Brown, L. D., Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, *24*(6), 2384–2398.

Carroll, R. J., Van Rooij, A. C. M., Ruymgaart, F. H. (1991). Theoretical aspects of ill-posed problems in statistics. *Acta Applicandae Mathematica*, *24*(2), 113–140.

Claeskens, G., Krivobokova, T., Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, *96*(3), 529–544.

Cox, D. D. (1983). Asymptotics for *M*-type smoothing splines. *The Annals of Statistics*, *11*(2), 530–551.

Cox, D. D. (1988). Approximation of method of regularization estimators. *The Annals of Statistics*, *16*(2), 694–712.

Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, *31*(4), 377–403.

Dudley, R. M. (2002). *Real analysis and probability*. Cambridge: Cambridge University Press.

Eilers, P. H. C., Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–121.

Feinberg, E. A., Kasyanov, P. O., Zadoianchuk, N. V. (2014). Fatou's lemma for weakly converging probabilities. *Theory of Probability & Its Applications*, *58*(4), 683–689.

Goldenshluger, A., Pereverzev, S. V. (2000). Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probability Theory and Related Fields*, *118*(2), 169–186.

Hall, P., Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, *92*(1), 105–118.

Hastie, T. J., Tibshirani, R. J. (1990). *Generalized additive models*. Boca Raton: Chapman and Hall.

Hurvich, C. M., Simonoff, J. S., Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *60*(2), 271–293.

Kauermann, G., Krivobokova, T., Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *71*(2), 487–503.

Kimeldorf, G. S., Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, *41*(2), 495–502.

Kimeldorf, G. S., Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, *33*(1), 82–95.

Kou, S. C., Efron, B. (2002). Smoothers and the C_p, generalized maximum likelihood, and extended exponential criteria: A geometric approach. *Journal of the American Statistical Association*, *97*(459), 766–782.

Lai, M.-J., Wang, L. (2013). Bivariate penalized splines for regression. *Statistica Sinica*, *23*, 1399–1417.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, *15*(3), 958–975.

Li, Y., Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, *95*(2), 415–436.

Lukas, M. A. (2006). Robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems*, *22*(5), 1883–1902.

Mair, B. A., Ruymgaart, F. H. (1996). Statistical inverse estimation in Hilbert scales. *SIAM Journal on Applied Mathematics*, *56*(5), 1424–1444.

Mallows, C. L. (1973). Some comments on C_P. *Technometrics*, *15*(4), 661–675.

Nychka, D. W., Cox, D. D. (1989). Convergence rates for regularized solutions of integral equations from discrete noisy data. *The Annals of Statistics*, *17*(2), 556–572.

Ragozin, D. L. (1983). Error bounds for derivative estimates based on spline smoothing of exact or noisy data. *Journal of Approximation Theory*, *37*(4), 335–355.

Sakamoto, Y., Ishiguro, M., Kitagawa, G. (1986). *Akaike information criterion statistics*. Tokyo: KTK Scientific Publishers (KTK).

Shen, J., Wang, X. (2011). Estimation of monotone functions via P-splines: A constrained dynamical optimization approach. *SIAM Journal on Control and Optimization*, *49*(2), 646–671.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, *12*(3), 898–916.

Speckman, P. L. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics*, *13*(3), 970–983.

Speckman, P. L., Sun, D. (2001). Asymptotic properties of smoothing parameter selection in spline smoothing, Technical report, Department of Statistics, University of Missouri. http://www.stat.missouri.edu/~speckman/pub.html.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, *10*(4), 1040–1053.

Thorpe, M., Johansen, A. M. (2016). Convergence and rates for fixed-interval multiple-track smoothing using k-means type optimization. *Electronic Journal of Statistics*, *10*(2), 3693–3722.

Thorpe, M., Theil, F., Johansen, A. M., Cade, N. (2015). Convergence of the k-means minimization problem using Γ-convergence. *SIAM Journal on Applied Mathematics*, *75*(6), 2444–2474.

Utreras, F. I. (1981). Optimal smoothing of noisy data using spline functions. *SIAM Journal on Scientific and Statistical Computing*, *2*(3), 349–362.

Utreras, F. I. (1983). Natural spline functions, their associated eigenvalue problem. *Numerische Mathematik*, *42*(1), 107–117.

Utreras, F. I. (1985). Smoothing noisy data under monotonicity constraints existence, characterization and convergence rates. *Numerische Mathematik*, *47*(4), 611–625.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, *13*(4), 1378–1402.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

Wahba, G., Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross-validation. *Communications in Statistics*, *4*(1), 1–17.

Wand, M. P. (1999). On the optimal amount of smoothing in penalised spline regression. *Biometrika*, *86*(4), 936–940.

Wang, X., Shen, J., Ruppert, D. (2011). On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics*, *5*, 1–17.
Xiao, L., Li, Y., Apanasovich, T. V., Ruppert, D. (2012). Local asymptotics of P-splines. arXiv:1201.0708
Xiao, L., Li, Y., Ruppert, D. (2013). Fast bivariate P-splines: The sandwich smoother. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *75*(3), 577–599.
Yoshida, T., Naito, K. (2012). Asymptotics for penalized additive B-spline regression. *Journal of the Japan Statistical Society*, *42*(1), 81–107.
Yoshida, T., Naito, K. (2014). Asymptotics for penalised splines in generalised additive models. *Journal of Nonparametric Statistics*, *26*(2), 269–289.