A regression based smoothing spline Monte Carlo algorithm for pricing American options

Michael Kohler*

Department of Mathematics, University of Saarbrücken, Postfach 151150, 66041 Saarbrücken, Germany, email: kohler@math.uni-sb.de

October 11, 2006

Abstract

Pricing of American options can be achieved by solving optimal stopping problems. This in turn can be done by computing so-called continuation values, which we represent as regression functions defined recursively by using the continuation values of the next time step. We use Monte Carlo to generate data and apply smoothing spline regression estimates to estimate the continuation values from these data. All parameters of the estimate are chosen data dependent. Results concerning consistency and rate of convergence of the estimates are presented.

AMS classification: Primary 91B28, 60G40, 93E20; secondary 65C05, 93E24, 62G05.

Key words and phrases: American options, consistency, nonparametric regression, optimal stopping, rate of convergence, regression based Monte Carlo methods, smoothing spline.

1 Introduction

Many financial contracts allow for early exercise before expiry. Most of the exchange traded options contracts are of American type which allows the holder to choose any exercise date before expiry, or Bermudan with exercise dates restricted to a predefined discrete set of dates. Examples are mortages with embedded payment options or life insurance contracts which allow for early surrender. In this article we are interested in pricing such Bermuda options which can be considered as discrete time approximations of American options.

*Corresponding author. Tel: +49-681-302-2435, Fax: +49-681-302-6583

Running title: *Pricing American options*

To do this, we assume that the price V_0 of such an option is represented as solution of an optimal stopping problem

$$V_0 = \sup_{\tau \in \mathcal{T}(0,\ldots,T)} \mathbf{E}\left\{ f_\tau(X_\tau) \right\}. \tag{1}$$

Here f_t is the (discounted) payoff function, X_0, X_1, ..., X_T is the underlying stochastic process and $\mathcal{T}(0,\ldots,T)$ is the class of all $\{0,\ldots,T\}$-valued stopping times, i.e., $\tau \in \mathcal{T}(0,\ldots,T)$ is a measurable function of X_0, ..., X_T satisfying

$$\{\tau = k\} \in \mathcal{F}(X_0,\ldots,X_k) \quad \text{for all } k \in \{0,\ldots,T\}.$$

As a very simple example consider pricing of an American put option with strike K and initial stock value x_0. We assume that the stock value is modelled via Black Scholes theory by

$$X_t = x_0 \cdot \exp\left((r - 1/2 \cdot \sigma^2) \cdot t + \sigma \cdot W_t\right), \tag{2}$$

where $r > 0$ is the (given) discount rate, $\sigma > 0$ is the (given) volatility of the asset, x_0 is the initial stock price and $\{W_t : t \in \mathbb{R}_+\}$ is a Wiener process. If we sell the option at time $t > 0$ and the stock price is at this point x, we get the payoff

$$\max\{K - x, 0\},$$

and if we discount this payoff towards time zero, we get the discounted payoff function

$$f_t(x) = e^{-r \cdot t} \cdot \max\{K - x, 0\}. \tag{3}$$

But even if all the parameters are known (i.e., if x_0 and K are given and if we estimate the volatility σ and the discount rate from observed data from the past), it is not obvious how we can compute the price

$$V_0 = \sup_{\tau \in \mathcal{T}(0,\ldots,T)} \mathbf{E}\left\{ e^{-r \cdot \tau} \cdot \max\{K - X_\tau, 0\} \right\}$$

of the corresponding American option. The purpose of this article is to develop an algorithm which is able to compute an approximation of the price (1) even in case that the stock price is not modelled by a simple Black Scholes model as in (2) and that the payoff function is not as simple as in (3). In particular the method of this article is also applicable in case that the process X_t is adjusted to observed data by time series estimation as described, e.g., in Franke and Diagne (2002), or that the payoff function is based on several underlyings, e.g., on the maximum of several stock values.

In the sequel we assume that X_0, X_1, $\ldots$, X_T is a $[-A, A]^d$–valued Markov process recording all necessary information about financial variables including prices of the underlying assets as well as additional risk factors driving stochastic volatility or stochastic interest rates. Neither the Markov property nor the form of the payoff as a function of the state X_t is restrictive and can always be achieved by including supplementary variables. But usually in modelling financial processes one models them by unbounded processes. In this case we choose a large value $A > 0$ and replace X_t by its bounded approximation

$$X_t^A = X_{\min\{t,\tau_A\}} \quad \text{where} \quad \tau_A = \inf\{s \geq 0 \,:\, X_s \notin [-A, A]^d\}.$$

(Here we assume for simplicity that the stochastic process has continuous paths in order to be able to neglect an additional truncation of X_t^A). This boundedness assumption enables us to estimate the price of the American option from samples of polynomial size in the number of free parameters, in contrast to Monte Carlo estimation from standard (unbounded) Black Scholes models, where Glasserman and Yu (2004) showed that samples of exponential size in the number of free parameters are needed.

The computation of (1) can be done by determination of an optimal stopping rule $\tau^* \in \mathcal{T}(0, \ldots, T)$ satisfying

$$V_0 = \mathbf{E}\{f_{\tau^*}(X_{\tau^*})\}. \tag{4}$$

Let

$$q_t(x) = \sup_{\tau \in \mathcal{T}(t+1,\ldots,T)} \mathbf{E}\left\{f_\tau(X_\tau)|X_t = x\right\} \tag{5}$$

be the so–called continuation value describing the value of the option at time t given $X_t = x$ and subject to the constraint of holding the option at time t rather than exercising it. Here $\mathcal{T}(t+1, \ldots, T)$ is the class of all $\{t+1, \ldots, T\}$–valued stopping times. It can be shown that

$$\tau^* = \inf\{s \geq 0 \,:\, q_s(X_s) \leq f_s(X_s)\} \tag{6}$$

satisfies (4), i.e., τ^* is an optimal stopping time (cf., e.g., Chow, Robbins and Siegmund (1971) or Shiryayev (1978)). Therefore it suffices to compute the continuation values (5) in order to solve the optimal stopping problem (1).

The continuation values satisfy the dynamic programming equations

$$\begin{aligned} q_T(x) &= 0, \\ q_t(x) &= \mathbf{E}\left\{\max\{f_{t+1}(X_{t+1}), q_{t+1}(X_{t+1})\}|X_t = x\right\} \quad (t = 0, 1, \ldots, T-1). \end{aligned} \tag{7}$$

Indeed, analogously to (6) we have

$$q_t(x) = \mathbf{E}\{f_{\tau_t^*}(X_{\tau_t^*})|X_t = x\} \quad \text{where} \quad \tau_t^* = \inf\{s \geq t+1 | q_s(X_s) \leq f_s(X_s)\},$$

hence by using the Markov property of $\{X_s\}_{s=0,\ldots,T}$ we get

$$\begin{aligned} & q_t(X_t) \\ & = \mathbf{E}\left\{f_{t+1}(X_{t+1}) \cdot I_{\{q_{t+1}(X_{t+1}) \leq f_{t+1}(X_{t+1})\}} + f_{\tau_{t+1}^*}(X_{\tau_{t+1}^*}) \cdot I_{\{q_{t+1}(X_{t+1}) > f_{t+1}(X_{t+1})\}} | X_t\right\} \\ & = \mathbf{E}\{\mathbf{E}\{\ldots|X_0, \ldots, X_{t+1}\}|X_0, \ldots, X_t\} \\ & = \mathbf{E}\left\{\max\{f_{t+1}(X_{t+1}), q_{t+1}(X_{t+1})\}|X_t\right\}. \end{aligned}$$

Unfortunately, the conditional expectation in (7) can in general not be computed in applications. The basic idea of regression based Monte Carlo methods for pricing American options is to apply recursively regression estimates to artificially created samples of

$$(X_t, \max\{f_{t+1}(X_{t+1}), \hat{q}_{t+1}(X_{t+1})\})$$

(so–called Monte Carlo samples) to construct estimates $\hat{q}_t$ of q_t. In connection with linear regression this was proposed in Tsitsiklis and Van Roy (1999), and, based on a different regression estimation than (7), in Longstaff and Schwartz (2001). Modern nonparametric least squares regression estimates have been applied and investigated in this context in Egloff (2005) and Egloff, Kohler and Todorovic (2006).

In this article we propose to use smoothing spline regression estimates in order to compute the conditional expectation in (7). This is in particular very promising since smoothing spline estimates are able to extend gradually the form of the estimate from completely parametric linear models to fully nonparametric models. And given the simple form of the payoff function in most applications (cf., e.g., (3)) it is not clear whether in applications the regression functions will be sufficient complicated that a fully nonparametric estimate (like the ones used in Egloff (2005) and Egloff, Kohler and Todorovic (2006)) is really necessary. Here the smoothing spline estimate with data-driven choice of the smoothing parameter seems to be a promising compromise between very simple linear regression and complicated nonparametric regression estimates.

Below we define smoothing spline estimates of the continuation values where all parameters of the estimates are chosen using the given data only. We will show that these estimates are universally consistent in the sense that their L_2 errors converges to zero in probability for all distributions. Furthermore, under regularity conditions on the smoothness of the continuation values we will analyze the rate of convergence of the estimates.

The precise definition of the estimates and the main results concerning consistency and rate of convergence of the estimate will be described in Section 2. The proofs will be given in Section 3.

2 Main results

2.1 Smoothing spline regression estimates

In this subsection we explain how the smoothing spline estimate of a regression function $m(x) = \mathbf{E}\{Y|X=x\}$ given a sample

$$\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \tag{8}$$

of the distribution of a $[-A, A]^d \times \mathbb{R}$–valued random variable (X, Y) is defined.

Let $k \in \mathbb{N}$ with $2k > d$. Denote by $W^k([-A, A]^d)$ the Sobolev space

$$\left\{ f \; : \; \frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} \in L_2([-A, A]^d) \text{ for all } \alpha_1, \ldots, \alpha_d \in \mathbb{N} \text{ with } \alpha_1 + \ldots + \alpha_d = k \right\}.$$

The condition $2k > d$ implies that the functions in $W^k([-A, A]^d)$ are continuous and hence the value of a function at a point is well defined. Set

$$J_k^2(f) = \sum_{\alpha_1, \ldots, \alpha_d \in \mathbb{N}, \, \alpha_1 + \ldots + \alpha_d = k} \frac{k!}{\alpha_1! \cdot \ldots \cdot \alpha_d!} \int_{\mathbb{R}^d} \left| \frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) \right|^2 dx.$$

Let $\lambda \in \mathbb{R}_+$. The smoothing spline estimate $\tilde{m}_{n,(k,\lambda)}$ of a regression function is defined by

$$\tilde{m}_{n,(k,\lambda)}(\cdot) = \arg \min_{f \in W^k([-A,A]^d)} \left(\frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 + \lambda \cdot J_k^2(f) \right). \tag{9}$$

Observe that $\tilde{m}_{n,(k,\lambda)}$ depends on the data (8) and that we have suppressed this in our notation.

Let $l = \binom{d+k-1}{d}$ and let $\phi_1, \ldots, \phi_l$ be all monomials $x_1^{\alpha_1} \cdot \ldots \cdot x_d^{\alpha_d}$ of total degree $\alpha_1 + \ldots + \alpha_d$ less than k. Define $R : \mathbb{R}_+ \to \mathbb{R}$ by

$$R(u) = \begin{cases} u^{2k-d} \cdot \log(u) & \text{if} \quad 2k-d \text{ is even,} \\ u^{2k-d} & \text{if} \quad 2k-d \text{ is odd,} \end{cases}$$

and denote the euclidean norm of a vector $x \in \mathbb{R}^d$ by $\|x\|_2$. It follows from Section V in Duchon (1976) that there exists a function of the form

$$\tilde{m}_{n,(k,\lambda)}(x) = \sum_{i=1}^{n} a_i R(\|x - X_i\|_2) + \sum_{j=1}^{l} b_j \phi_j(x) \tag{10}$$

which achieves the minimum in (9), and that the coefficients $a_1, \ldots, a_n, b_1, \ldots, b_l \in \mathbb{R}$ of this function can be computed by solving a linear system of equations. Under some additional assumptions on the $X_1, \ldots, X_n$ this is also shown in Section 2.4 of Wahba (1990).

The parameter $\lambda \in \mathbb{R}_+$ is a smoothing parameter of the estimate which controls how much the data is smoothed. For $\lambda = 0$ the estimate will interpolate the given data (provided the x-values of the sample are pairwise distinct), while for λ large, i.e., for $\lambda \to \infty$, the estimate will be a polynomial of degree $k-1$ fitted to the data by least squares. In this sense the choice of the smoothing parameter enables to adjust the estimate between a parametric estimate ($\lambda \approx \infty$) and a completely nonparametric estimate (for small λ).

2.2 Smoothing spline estimates for pricing of American options

Let $X_0, X_1, \ldots, X_T$ be a $[-A, A]^d$–valued Markov process and let f_t be the discounted payoff function which we assume to be bounded in absolute value by L. In the sequel we describe an algorithm to estimate the continuation values q_t (defined by (5)) recursively. To do this we generate artificial independent Markov processes $\{X_{i,t}^{(l)}\}_{t=0,\ldots,T}$ ($l = 0, 1, \ldots, T, i = 1, 2, \ldots, n$) which are identically distributed as $\{X_t\}_{t=0,\ldots,T}$. Then we use these so-called Monte Carlo samples to generate recursively data to estimate q_t by using the regression representation given in (7).

We start with

$$q_{n,T}(x) = 0 \quad (x \in \mathbb{R}^d).$$

Given an estimate $q_{n,t+1}$ of q_{t+1}, we estimate

$$q_t(x) = \mathbf{E}\left\{\max\{f_{t+1}(X_{t+1}), q_{t+1}(X_{t+1})\}|X_t = x\right\}$$

by applying a smoothing spline regression estimate to an "approximative" sample of

$$\left(X_t, \max\{f_{t+1}(X_{t+1}), q_{t+1}(X_{t+1})\}\right).$$

With the notation

$$\hat{Y}_{i,t}^{(t)} = \max\{f_{t+1}(X_{i,t+1}^{(t)}), q_{n,t+1}(X_{i,t+1}^{(t)})\}$$

(where we have suppressed the dependency of $\hat{Y}_{i,t}^{(t)}$ on n) this "approximative" sample is given by

$$\left\{\left(X_t^{(t)}, \hat{Y}_{i,t}^{(t)}\right) \quad : \quad i = 1, \ldots, n\right\}. \tag{11}$$

Observe that this sample depends on the t-th sample of $\{X_s\}_{s=0,\ldots,T}$ and $q_{n,t+1}$, i.e., for each time step t we use a new sample of the stochastic process $\{X_s\}_{s=0,\ldots,T}$ in order to define our data (11).

To choose the smoothing parameters k and λ of the smoothing spline regression estimate fully automatically we use splitting of the sample. I.e., we subdivide the data (11) in a learning sample of size $n_l = \lceil n/2 \rceil$ and a testing sample of size $n_t = n - n_l$ and define for given $\lambda \in \mathbb{R}_+$ and $k \in \mathbb{N}_0$ a regression estimate of q_t by

$$\tilde{q}_{n_l,t}^{(k,\lambda)}(\cdot) = \arg \min_{f \in W^k([-A,A]^d)} \left(\frac{1}{n_l} \sum_{i=1}^{n_l} |f(X_{i,t}^{(t)}) - \hat{Y}_{i,t}^{(t)}|^2 + \lambda \cdot J_k^2(f)\right)$$

and

$$q_{n_l,t}^{(k,\lambda)}(x) = T_L \tilde{q}_{n_l,t}^{(k,\lambda)}(x) \quad (x \in \mathbb{R}^d),$$

where $T_L z = \max\{-L, \min\{L, z\}\}$ for $z \in \mathbb{R}$. Then we minimize the empirical L_2 error on the discrete parameter set

$$\mathcal{P}_n = \left\{(k, \lambda) \quad : \quad \lambda = \frac{i}{n} \text{ for some } i \in \{0, 1, \ldots, n^2\}, k \in \left\{\lceil \frac{d}{2} \rceil, \ldots, K\right\}\right\}$$

(where $K \geq \lceil d/2 \rceil$ is a given natural number) in order to choose the value of the parameter. I.e., we choose

$$(\lambda_t^*, k_t^*) = \arg \min_{(k,\lambda) \in \mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n_l,t}^{(k,\lambda)}(X_{i,t}^{(t)}) - \hat{Y}_{i,t}^{(t)}|^2$$

and define our final estimate of q_t by

$$q_{n,t}(x) = q_{n_l,t}^{(\lambda_t^*, k_t^*)}(x) \quad (x \in \mathbb{R}^d).$$

2.3 Consistency and rate of convergence of the estimates

Our main results are the following two theorems concerning consistency and rate of convergence of the estimates.

Theorem 1 *Let $A, L > 0$. Assume that $X_0, X_1, \ldots, X_T$ is a $[-A, A]^d$–valued Markov process and that the discounted payoff function f_t is bounded in absolute value by L. Then the estimates $q_{n,t}$ defined in the previous subsections satisfy for any $t \in \{0, 1, \ldots, T\}$*

$$\int |q_{n,t}(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx) \to 0 \quad \textit{in probability.}$$

The above theorem shows that the L_2 error of our estimates converges to zero in probability for sample size (of the Monte Carlo sample) tending to infinity. In view of an application with necessarily finite sample size it would be nice to know how quickly the error converges to zero for sample size tending to infinity. It is well-known that assumptions on the underlying distribution, in particular on the smoothness of the regression function, are necessary in order to be able to derive non-trivial rate of convergence results in nonparametric regression (see, e.g., Cover (1968), Devroye (1982) or Chapter 3 in Györfi et al. (2002)). In the next theorem we assume $q_t \in W^k([-A, A]^d)$ and show that under this assumption our estimates achieve (up to some logarithmic factor) the corresponding optimal rate of convergence. Here we will write $U_n = O_{\mathbf{P}}(V_n)$ for random variables U_n and nonnegative random variables V_n, if

$$\mathbf{P}\{|U_n| > const \cdot V_n\} \to 0 \quad (n \to \infty)$$

for some constant $const > 0$.

Theorem 2 *Let $A, L > 0$. Assume that $X_0, X_1, \ldots, X_T$ is a $[-A, A]^d$–valued Markov process, that the discounted payoff function f_t is bounded in absolute value by L and that the continuation values satisfy*

$$q_t \in \{f \in W^{k^*}([-A, A]^d) : J^2_{k^*}(f) \le C\} \quad (t = 0, 1, \ldots, T-1)$$

for some $k^* \in \{\lceil d/2 \rceil, \ldots, K\}$ and some $C > 0$. Then the estimates $q_{n,t}$ defined in the previous subsections satisfy for any $t \in \{0, 1, \ldots, T\}$

$$\int |q_{n,t}(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx) = O_{\mathbf{P}}\left(C^{\frac{d}{2k^*+d}} \cdot \left(\frac{\log(n)}{n} \right)^{\frac{2k^*}{2k^*+d}} \right).$$

Remark 1. It follows from Stone (1982) that the rate of convergence in Theorem 2 is optimal (in some Minimax sense) up to a logarithmic factor.

Remark 2. The definition of the estimate in Theorem 2 does not depend on the smoothness of the continuation values measured by k and $J_k(q_t)$, nevertheless the estimate is able to achieve (up to some logarithmic factor) the corresponding optimal rate of convergence. In this sense the estimate is able to adapt automatically to the smoothness of the continuation values.

Remark 3. It follows from the proof of the above theorems, that the assumption that f_t is bounded in absolute value by L can be replaced by $|f_t(X_t)| \leq L$ $a.s.$ Since we assume that the Markov process $X_0, \ldots, X_T$ is bounded, this in turn is satisfied for more or less all payoff functions occuring in practice.

Remark 4. The estimate $q_{n,T}, q_{n,T-1}, \ldots, q_{n,0}$ can be used to approximate the optimal stopping time τ^* defined in (4) by

$$\hat{\tau} = \inf \{s \geq 0 \quad : \quad q_{n,s}(X_s) \leq f_s(X_s)\}$$

and to estimate the price V_0 of the American option defined in (1) by a Monte Carlo estimate of $\mathbf{E}\{f_{\hat{\tau}}(X_{\hat{\tau}})\}$. In case of $X_0 = x_0$ $a.s$ (with initial stock value x_0) the price V_0 of the American option can also be estimated by

$$\max\{f_0(x_0), q_{n,0}(x_0)\}$$

and it is easy to see that the above theorems imply consistency and corresponding rate of convergence results of this estimate.

2.4 Discussion of related results

Smoothing spline regression estimates have been studied by many authors, see, e.g., the monographs by Eubank (1988) and Wahba (1990) and the literature cited therein. In the context of random design regression consistency and rate of convergence of univariate smoothing spline regression estimates have been studied by means of empirical process theory by van de Geer (1987, 1988, 1990). Kohler and Krzyżak (2001) proved that suitable defined smoothing spline regression estimates are universally consistent.

In the proofs we analyze the problem of pricing American options in the context of regression estimation with additional measurement errors in the dependent data (cf.

Section 3). The idea to consider regression estimation with additional measurement errors in the dependent data was proposed in Kohler (2006). There results concerning the rate of convergence of least squares estimates have been derived and have been used to analyze the rate of convergence of regression estimates based on censored data. Kohler, Kul and Mathé (2004) studied adaptive least squares estimates based on splitting of the sample in this context, and proved also an additional result concerning universal consistency of the estimates. Dippon and Winter (2006) studied smoothing spline estimates for regression estimation with additional measurement errors in the dependent data. In particular, they extended the consistency result of Kohler and Krzyżak (2001) to this context and analyzed the rate of convergence of the estimates. The current article shows how to apply these results to the problem of pricing American options.

Various Monte Carlo methods in financial engineering are described in detail in the monography Glasserman (2004). Regression based Monte Carlo methods for pricing American options have been proposed by Tsitsiklis and Van Roy (1999) and Longstaff and Schwartz (2001) based on different regression representations of the continuation values. Egloff (2005) presents a generalization of both regression representations. The first two articles use linear regression (i.e., a standard parametric approach) to estimate the regression function. In contrast, Egloff (2005) focuses on nonparametric least squares estimates in this context. Egloff, Kohler and Todorovic (2006) modify the estimates of Egloff (2005) such that they are easier to compute in an application and such that all parameters of the estimates are chosen only by aid of the given data. In both papers results concerning consistency and rate of convergence of suitably defined least squares estimates are proven.

In contrast to the above papers, the current paper focuses on smoothing spline estimates. Since these estimates are able to extend gradually the form of the estimate from completely parametric linear models to fully nonparametric models they seemed to be in particular promising for pricing of American options, where usually the payoff function is so simple that it is not clear whether the continuation values are sufficient complicated that a fully nonparametric regression estimate is really useful.

3 Proofs

3.1 Regression estimation in case of additional measurement errors in the dependent variable

In the proof we will use results concerning regression estimation in case of additional measurement errors in the dependent variable, which we describe in the sequel.

Let $(X, Y), (X_1, Y_1), \dots$ be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ valued random variables with $\mathbf{E}Y^2 < \infty$. Let $m(x) = \mathbf{E}\{Y|X = x\}$ be the corresponding regression function. Assume that we want to estimate m from observed data, but instead of a sample

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

of (X, Y) we have only available a set of data

$$\bar{\mathcal{D}}_n = \{(X_1, \bar{Y}_{1,n}), \dots, (X_n, \bar{Y}_{n,n})\}$$

where the only assumption on $\bar{Y}_{1,n}, \dots, \bar{Y}_{n,n}$ is that the measurement error

$$\frac{1}{n} \sum_{i=1}^{n} |Y_i - \bar{Y}_{i,n}|^2 \tag{12}$$

is small. In particular we do not assume that the random variables in $\bar{\mathcal{D}}_n$ are independent or identically distributed. In the sequel we are interested in the influence of the measurement error (12) on the L_2 error of a regression estimate applied to the data $\bar{\mathcal{D}}_n$.

As we do not assume anything on the difference between the true y-values Y_i and the observed values $\bar{Y}_{i,n}$ besides the assumption that (12) is small, it is clear that there is no chance to get rid of this measurement error completely. But a natural conjecture is that a small measurement error (12) does only slightly influence the L_2 error of suitably defined regression estimates. That this conjecture is indeed true was proven for least squares estimates in Kohler (2006) and for smoothing spline estimates in Dippon and Winter (2006). We describe next the part of this result, which we will need in the proofs of our main results.

Assume

$$X_i \in [-A, A]^d \text{ a.s. and } Y_i, \bar{Y}_{i,n} \in [-L, L] \text{ a.s.}$$

$(i = 1, \ldots, n)$ and define for $k \in \mathbb{N}_0$ and $\lambda_n > 0$ the smoothing spline estimate $m_{n,(k,\lambda_n)}$ by

$$\tilde{m}_{n,(k,\lambda_n)}(\cdot) = \arg \min_{f \in W^k([-A,A]^d)} \left(\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - \bar{Y}_{i,n}|^2 + \lambda_n \cdot J_k^2(f) \right)$$

and

$$m_{n,(k,\lambda_n)}(x) = T_L \tilde{m}_{n,(k,\lambda_n)}(x) \quad (x \in \mathbb{R}^d)$$

(where $T_L z = \max\{-L, \min\{L, z\}\}$ for $z \in \mathbb{R}$). Then the following result holds.

Lemma 1 *Assume $2k > d$. Under the above assumptions we have*

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} |m_{n,(k,\lambda_n)}(X_i) - m(X_i)|^2 + \lambda_n \cdot J_k^2(\tilde{m}_{n,(k,\lambda_n)}) \\ &= O_{\mathbf{P}} \left(\frac{1}{n} \sum_{i=1}^{n} |Y_i - \bar{Y}_{i,n}|^2 + \frac{\log n}{n} \cdot \lambda_n^{-d/2k} + \lambda_n \cdot J_k^2(m) + \frac{\log n}{n} \right). \end{aligned}$$

Proof: The result follows from the proof Lemma A.1 in Dippon and Winter (2006). For the sake of completeness we give a (short) outline of the proof. The crucial step is to extend Lemma 1 in Kohler (2006) in order to show that

$$\begin{aligned} &\tfrac{1}{n} \sum_{i=1}^{n} |m_{n,(k,\lambda_n)}(X_i) - m(X_i)|^2 + \lambda_n \cdot J_k^2(\tilde{m}_{n,(k,\lambda_n)}) \\ &> t + \tfrac{64}{n} \sum_{i=1}^{n} |Y_i - \bar{Y}_{i,n}|^2 + 2 \cdot \lambda_n \cdot J_k^2(m) \end{aligned}$$

implies

$$\begin{aligned} t &\leq \frac{1}{n} \sum_{i=1}^{n} |m_{n,(k,\lambda_n)}(X_i) - m(X_i)|^2 + \lambda_n \cdot J_k^2(\tilde{m}_{n,(k,\lambda_n)}) \\ &\leq \frac{8}{n} \sum_{i=1}^{n} (m_{n,(k,\lambda_n)}(X_i) - m(X_i)) \cdot (Y_i - m(X_i)). \end{aligned}$$

Hence

$$\begin{aligned} &\mathbf{P}\Bigg\{ \frac{1}{n} \sum_{i=1}^{n} |m_{n,(k,\lambda_n)}(X_i) - m(X_i)|^2 + \lambda_n \cdot J_k^2(\tilde{m}_{n,(k,\lambda_n)}) \\ &\qquad\qquad > t + \frac{64}{n} \sum_{i=1}^{n} |Y_i - \bar{Y}_{i,n}|^2 + 2 \cdot \lambda_n \cdot J_k^2(m) \Bigg\} \\ &\leq \sum_{j=0}^{\infty} \mathbf{P}\Big\{ 2^j t \cdot I_{\{j \neq 0\}} \leq \lambda_n \cdot J_k^2(\tilde{m}_{n,(k,\lambda_n)}) < 2^{j+1} t, \ldots \Big\} \\ &\leq \sum_{j=0}^{\infty} \mathbf{P}\Bigg\{ J_k^2(\tilde{m}_{n,(k,\lambda_n)}) < 2^{j+1} t / \lambda_n, 2^j t \leq \frac{8}{n} \sum_{i=1}^{n} (m_{n,(k,\lambda_n)}(X_i) - m(X_i)) \cdot (Y_i - m(X_i) \Bigg\}. \end{aligned}$$

The last probability is then bounded by Corollary 8.3 in van de Geer (2000) using the bound on the covering number given in Lemma 3 below. □

The above lemma enables us to analyze the rate of convergence of the estimate for fixed k and λ_n. Next we explain how we can use the data to choose appropriate values for k and λ_n. To do this we use splitting of the sample with a learning sample

$$\bar{\mathcal{D}}_{n_l} = \left\{(X_1, \bar{Y}_{1,n}), \dots, (X_{n_l}, \bar{Y}_{n_l,n})\right\}$$

of size $n_l = \lceil n/2 \rceil$ and a testing sample

$$\left\{(X_{n_l+1}, \bar{Y}_{n_l+1,n}), \dots, (X_n, \bar{Y}_{n,n})\right\}$$

of size $n_t = n - n_l$. For fixed k and $\lambda > 0$ we use the learning sample to define a smoothing spline estimate $m_{n_l,(k,\lambda)}$ by

$$\tilde{m}_{n_l,(k,\lambda)}(\cdot) = \arg \min_{f \in W^k([-A,A]^d)} \left(\frac{1}{n_l} \sum_{i=1}^{n_l} |f(X_i) - \bar{Y}_{i,n}|^2 + \lambda \cdot J_k^2(f) \right)$$

and

$$m_{n_l,(k,\lambda)}(x) = T_L \tilde{m}_{n_l,(k,\lambda)}(x) \quad (x \in \mathbb{R}^d).$$

Let $\mathcal{P}_n$ be a finite set of parameters (k, λ). Next we choose $(\hat{k}, \hat{\lambda}) \in \mathcal{P}_n$ by minimizing the empirical L_2 risk on the testing sample, i.e., we set

$$m_n(x) = m_{n_l,(\hat{k},\hat{\lambda})}(x) \quad (x \in \mathbb{R}^d),$$

where

$$(\hat{k}, \hat{\lambda}) = \arg \min_{(k,\lambda) \in \mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |m_{n_l,(k,\lambda)}(X_i) - \bar{Y}_{i,n}|^2.$$

Then the following result holds.

Lemma 2 *Assume* $|\mathcal{P}_n| \to \infty$ $(n \to \infty)$. *Then*

$$\frac{1}{n_t} \sum_{i=n_l+1}^{n} |m_n(X_i) - m(X_i)|^2$$
$$= O_{\mathbf{P}} \left(\frac{\log |\mathcal{P}_n|}{n_t} + \frac{1}{n_t} \sum_{i=n_l+1}^{n} |Y_i - \bar{Y}_{i,n}|^2 + \min_{(k,\lambda) \in \mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |m_{n_l,(k,\lambda)}(X_i) - m(X_i)|^2 \right).$$

Proof. See Lemma 4 in Kohler, Kul and Mathé (2004). □

3.2 Auxiliary results

In this subsection we formulate two auxiliary results which we will use in the proof of our main results. To be able to do this we need the notion of covering numbers.

Definition 1 *Let $q \geq 1$, $l \in \mathbb{N}$ and let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^l \to \mathbb{R}$. The covering number $\mathcal{N}_q(\epsilon, \mathcal{F}, x_1^n)$ is defined for any $\epsilon > 0$ and $x_1^n = (x_1, ..., x_n) \in (\mathbb{R}^l)^n$ as the smallest integer k such that there exist functions $g_1, ..., g_k : \mathbb{R}^l \to \mathbb{R}$ with*

$$\min_{1 \leq i \leq k} \left(\frac{1}{n} \sum_{j=1}^{n} |f(x_j) - g_i(x_j)|^q \right)^{1/q} \leq \epsilon$$

for each $f \in \mathcal{F}$.

In the proofs of our main results we will need the following two lemmas.

Lemma 3 *(Lemma 3 in Kohler, Krzyżak and Schäfer (2002)).*
Let $A, L, c > 0$ and set

$$\mathcal{F} = \left\{ T_L f \ : \ f \in W^k([-A, A]^d) \text{ and } J_k^2(f) \leq c \right\}.$$

Then there exists a constant $c_d \in \mathbb{R}_+$ depending only on d and A such that for any $\epsilon > 0$ and all $x_1, \ldots, x_n \in [-A, A]^d$

$$\log \mathcal{N}_2(\epsilon, \mathcal{F}, x_1^n) \leq c_d(k^d + 1) \left(\left(\frac{\sqrt{c}}{\epsilon} \right)^{d/k} + 1 \right) \cdot \log \left(\frac{64 e L^2 n}{\epsilon^2} \right) \cdot I_{\{\epsilon \leq L\}}. \tag{13}$$

Lemma 4 *(Lemma 5 in Kohler (2006)).*
Let $L \geq 1$, let $m : \mathbb{R}^d \to [-L, L]$ and let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [-L, L]$. Let $0 < \epsilon < 1$ and $\alpha > 0$. Assume that

$$\sqrt{n} \epsilon \sqrt{\alpha} \geq 1152 L$$

and that, for all $x_1, \ldots, x_n \in \mathbb{R}^d$ and all $\delta \geq 2L^2\alpha$,

$$\frac{\sqrt{n}\epsilon\delta}{768\sqrt{2}L^2} \geq$$
$$\int_{\frac{\epsilon\delta}{128L^2}}^{\sqrt{\delta}} \left(\log \mathcal{N}_2 \left(\frac{u}{4L}, \left\{ f - m \ : \ f \in \mathcal{F}, \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - m(x_i)|^2 \leq \frac{\delta}{L^2} \right\}, x_1^n \right) \right)^{1/2} du.$$

Then

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}} \frac{\left|\mathbf{E}\{|f(X)-m(X)|^2\} - \frac{1}{n}\sum_{i=1}^n |f(X_i)-m(X_i)|^2\right|}{\alpha + \mathbf{E}\{|f(X)-m(X)|^2\} + \frac{1}{n}\sum_{i=1}^n |f(X_i)-m(X_i)|^2} > \epsilon\right\}$$
$$\leq 15\exp\left(-\frac{n\alpha\epsilon^2}{512\cdot 2304 L^2}\right).$$

3.3 Proof of Theorem 1

In the sequel we will show

$$\int |q_{n,s}(x) - q_s(x)|^2 \mathbf{P}_{X_s}(dx) \to 0 \quad \text{in probability} \tag{14}$$

for all $s \in \{0, 1, \ldots, T\}$.

For $s = T$ we have $q_{n,T}(x) = 0 = q_T(x)$, so the assertion is trivial. So let $t < T$ and assume that the assertion holds for $s \in \{t+1, \ldots, T\}$. By induction it suffices to show (14) for $s = t$, which we will show in the sequel in nine steps.

In the first step of the proof we show

$$\int |q_{n,t}(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx) - \frac{1}{n_t}\sum_{i=n_l+1}^{n} |q_{n,t}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 \to 0 \quad \text{in probability.}$$

Let $\mathcal{D}_{n,t}$ be the set of all $X_{j,s}^{(r)}$ with either $r \geq t+1, s \in \{0, \ldots, T\}$ and $j \in \{1, \ldots, n\}$ or $r = t, s = t$ and $j \in \{1, \ldots, n_l\}$. Conditioned on $\mathcal{D}_{n,t}$,

$$\{q_{n_l,t}^{(k,\lambda)} : (k, \lambda) \in \mathcal{P}_n\}$$

consists of $|\mathcal{P}_n|$ different functions. Using the boundedness of $q_{n_l,t}^{(k,\lambda)}$ and q_t by L we get by Hoeffding‘s inequality (cf., e.g., Lemma A.3 in Györfi et al. (2002)) for any $\epsilon > 0$:

$$\mathbf{P}\{\textstyle\int |q_{n,t}(x) - q_t(x)|^2\, \mathbf{P}_{X_t}(dx) - \frac{1}{n_t}\sum\limits_{i=n_l+1}^{n} |q_{n,t}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 > \epsilon\, |\mathcal{D}_{n,t}\}$$

$$\leq |\mathcal{P}_n| \max\nolimits_{(k,\lambda)\in\mathcal{P}_n} \mathbf{P}\{\textstyle\int |q_{n,t}^{(k,\lambda)}(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx)$$
$$\qquad\qquad -\textstyle\frac{1}{n_t}\sum\limits_{i=n_l+1}^{n} |q_{n,t}^{(k,\lambda)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 > \epsilon \mid \mathcal{D}_{n,t}\}$$

$$\leq |\mathcal{P}_n| \cdot \exp\left(-\tfrac{n_t\epsilon^2}{16L^4}\right)$$

$$= \exp\left(-n_t \cdot \left(\tfrac{\epsilon^2}{16L^4} - \tfrac{\log|\mathcal{P}_n|}{n_t}\right)\right)$$

$$\to 0 \qquad (n \to \infty),$$

since

$$\frac{\log(|\mathcal{P}_n|)}{n_t} \leq \frac{\log((n+1)^2 K)}{n/2-1} \to 0 \qquad (n \to \infty).$$

In the second step of the proof we show

$$\begin{aligned}
&\tfrac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n,t}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 \\
&= O_{\mathbf{P}}\Bigg(\tfrac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n,t+1}(X_{i,t+1}^{(t)}) - q_{t+1}(X_{i,t+1}^{(t)})|^2 + \tfrac{\log|\mathcal{P}_n|}{n_t} \\
&\qquad + \min_{(k,\lambda)\in\mathcal{P}_n} \tfrac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n_l,t}^{(k,\lambda)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 \Bigg).
\end{aligned}$$

To do this we use Lemma 2. In the context of Lemma 2 we have

$X_i = X_{i,t}^{(t)}, Y_i = \max\{f_{t+1}(X_{i,t+1}^{(t)}), q_{t+1}(X_{i,t+1}^{(t)})\}$ and $\bar{Y}_{i,n} = \max\{f_{t+1}(X_{i,t+1}^{(t)}), q_{n,t+1}(X_{i,t+1}^{(t)})\}$.

Observing

$$\frac{1}{n_t} \sum_{i=n_l+1}^{n} |Y_i - \bar{Y}_{i,n}|^2 \leq \frac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{t+1}(X_{i,t+1}^{(t)}) - q_{n,t+1}(X_{i,t+1}^{(t)})|^2$$

the assertion follows from Lemma 2 if we apply it conditioned on $\mathcal{D}_{n,t}$.

In the third step of the proof we observe

$$\tfrac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n,t+1}(X_{i,t+1}^{(t)}) - q_{t+1}(X_{i,t+1}^{(t)})|^2 - \int |q_{n,t+1}(x) - q_{t+1}(x)|^2 \mathbf{P}_{X_{t+1}}(dx) \to 0$$

in probability. Indeed, this follows as in the first step of the proof by an application of Hoeffding‘s inequality.

Choose $(k^*, \lambda_n) \in \mathcal{P}_n$ such that

$$\lambda_n \to 0 \quad (n \to \infty) \quad \text{and} \quad \frac{n \cdot \lambda_n^{d/(2k^*)}}{\log n} \to \infty \quad (n \to \infty).$$

In the fourth step of the proof we show that

$$\min_{(k,\lambda)\in\mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n_l,t}^{(k,\lambda)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 \to 0 \quad \text{in probability}$$

is implied by

$$\int |q_{n_l,t}^{(k^*,\lambda_n)}(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx) \to 0 \quad \text{in probability}.$$

To see this, we observe that we have as in the third step of the proof

$$\tfrac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n_l,t}^{(k^*,\lambda_n)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 - \int |q_{n_l,t}^{(k^*,\lambda_n)}(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx) \to 0$$

in probability, hence the assertion follows from

$$\min_{(k,\lambda)\in\mathcal{P}_n} \tfrac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n_l,t}^{(k,\lambda)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 \leq \tfrac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n_l,t}^{(k^*,\lambda_n)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2.$$

Given the results of the previous steps, we see that it suffices to show

$$\int |q_{n_l,t}^{(k^*,\lambda_n)}(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx) \to 0 \quad \text{in probability,}$$

which we do in the sequel by extending arguments in Kohler and Krzyżak (2001). The modification of these arguments was also described in the proof of Theorem 4.1 in Dippon and Winter (2006). For the sake of completeness we repeat the calculations here.

Let $\epsilon, \delta > 0$ and choose $q_{t,\epsilon} \in W^{k^*}([-A,A]^d)$ such that $q_{t,\epsilon}$ is bounded in supremum norm and such that

$$\int |q_t(x) - q_{t,\epsilon}(x)|^2 \mathbf{P}_{X_t}(dx) < \epsilon$$

(cf., e.g., Corollary A.1 in Györfi et al. (2002)). Set

$$Y_t = \max\{f_{t+1}(X_{t+1}), q_{t+1}(X_{t+1})\} \quad \text{and} \quad Y_{i,t}^{(t)} = \max\{f_{t+1}(X_{i,t+1}^{(t)}), q_{t+1}(X_{i,t+1}^{(t)})\}$$

and let $\bar{\mathcal{D}}_{n,t}$ be the set of all $X_{j,s}^{(r)}$ with $r \geq t$, $s \in \{0,\ldots,T\}$ and $j \in \{1,\ldots,n\}$. We will use the error decomposition

$$\begin{aligned}
&\int |q_{n_l,t}^{(k^*,\lambda_n)}(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx) \\
&= \mathbf{E}\left\{|q_{n_l,t}^{(k^*,\lambda_n)}(X_t) - Y_t|^2 |\bar{\mathcal{D}}_{n,t}\right\} - \mathbf{E}\left\{|q_t(X_t) - Y_t|^2\right\} \\
&= \mathbf{E}\left\{|q_{n_l,t}^{(k^*,\lambda_n)}(X_t) - Y_t|^2 |\bar{\mathcal{D}}_{n,t}\right\} - \frac{1}{n_l}\sum_{i=1}^{n_l} |q_{n_l,t}^{(k^*,\lambda_n)}(X_{i,t}^{(t)}) - Y_{i,t}^{(t)}|^2 \\
&\quad + \frac{1}{n_l}\sum_{i=1}^{n_l} |q_{n_l,t}^{(k^*,\lambda_n)}(X_{i,t}^{(t)}) - Y_{i,t}^{(t)}|^2 - (1+\delta)\cdot\frac{1}{n_l}\sum_{i=1}^{n_l} |q_{n_l,t}^{(k^*,\lambda_n)}(X_{i,t}^{(t)}) - \hat{Y}_{i,t}^{(t)}|^2 \\
&\quad + (1+\delta)\cdot\frac{1}{n_l}\sum_{i=1}^{n_l} |q_{n_l,t}^{(k^*,\lambda_n)}(X_{i,t}^{(t)}) - \hat{Y}_{i,t}^{(t)}|^2 - (1+\delta)\cdot\frac{1}{n_l}\sum_{i=1}^{n_l} |q_{t,\epsilon}(X_{i,t}^{(t)}) - \hat{Y}_{i,t}^{(t)}|^2 \\
&\quad + (1+\delta)\cdot\frac{1}{n_l}\sum_{i=1}^{n_l} |q_{t,\epsilon}(X_{i,t}^{(t)}) - \hat{Y}_{i,t}^{(t)}|^2 - (1+\delta)^2\cdot\frac{1}{n_l}\sum_{i=1}^{n_l} |q_{t,\epsilon}(X_{i,t}^{(t)}) - Y_{i,t}^{(t)}|^2 \\
&\quad + (1+\delta)^2\cdot\frac{1}{n_l}\sum_{i=1}^{n_l} |q_{t,\epsilon}(X_{i,t}^{(t)}) - Y_{i,t}^{(t)}|^2 - (1+\delta)^2\cdot\mathbf{E}\left\{|q_{t,\epsilon}(X_t) - Y_t|^2\right\} \\
&\quad + (1+\delta)^2\cdot\mathbf{E}\left\{|q_{t,\epsilon}(X_t) - Y_t|^2\right\} - \mathbf{E}\left\{|q_t(X_t) - Y_t|^2\right\} \\
&= \sum_{j=1}^{6} T_{j,n}.
\end{aligned}$$

In the fifth step of the proof we show

$$T_{1,n} \to 0 \quad \text{in probability.}$$

This follows as in Lemma 1 of Kohler and Krzyżak (2001), if we observe that Y_t is bounded in absolute value by L and that the definition of the estimate implies

$$\begin{aligned}
\lambda_n J_{k^*}^2(\tilde{q}_{n_l,t}^{(k^*,\lambda_n)}) &\leq \frac{1}{n_l}\sum_{i=1}^{n_l} |\tilde{q}_{n_l,t}^{(k^*,\lambda_n)}(X_{i,t}^{(t)}) - \hat{Y}_{i,t}^{(t)}|^2 + \lambda_n J_{k^*}^2(\tilde{q}_{n_l,t}^{(k^*,\lambda_n)}) \\
&\leq \frac{1}{n_l}\sum_{i=1}^{n_l} |0 - \hat{Y}_{i,t}^{(t)}|^2 + \lambda_n \cdot 0 \\
&\leq \frac{2}{n_l}\sum_{i=1}^{n_l} |\hat{Y}_{i,t}^{(t)} - Y_{i,t}^{(t)}|^2 + \frac{2}{n_l}\sum_{i=1}^{n_l} |Y_{i,t}^{(t)}|^2 \\
&\to 2 \cdot \mathbf{E} Y_t^2 \quad \text{in probability}
\end{aligned}$$

since

$$\frac{1}{n_l}\sum_{i=1}^{n_l} |\hat{Y}_{i,t}^{(t)} - Y_{i,t}^{(t)}|^2 \leq \frac{1}{n_l}\sum_{i=1}^{n_l} |q_{n,t+1}(X_{i,t}^{(t)}) - q_{t+1}(X_{i,t}^{(t)})|^2 \to 0 \quad \text{in probability} \tag{15}$$

by induction.

In the sixth step of the proof we show for every $\eta > 0$

$$\mathbf{P}\left\{\limsup_{n\to\infty} T_{j,n} > \eta\right\} = 0 \quad \text{for } j \in \{2, 4\}.$$

Using $(a+b)^2 \leq (1+\delta)a^2 + (1+1/\delta)b^2$ $(a, b > 0)$ we get

$$\begin{aligned}
T_{2,n} &= \frac{1}{n_l}\sum_{i=1}^{n_l} |q_{n_l,t}^{(k^*,\lambda_n)}(X_{i,t}^{(t)}) - \hat{Y}_{i,t}^{(t)} + \hat{Y}_{i,t}^{(t)} - Y_{i,t}^{(t)}|^2 - (1+\delta)\cdot \frac{1}{n_l}\sum_{i=1}^{n_l} |q_{n_l,t}^{(k^*,\lambda_n)}(X_{i,t}^{(t)}) - \hat{Y}_{i,t}^{(t)}|^2 \\
&\leq (1+\frac{1}{\delta}) \cdot \frac{1}{n_l}\sum_{i=1}^{n_l} |\hat{Y}_{i,t}^{(t)} - Y_{i,t}^{(t)}|^2 \\
&\to 0 \quad \text{in probability}
\end{aligned}$$

by (15). The result for $T_{4,n}$ follows in the same way.

In the seventh step of the proof we observe that we have by definition of the estimate

$$T_{3,n} \leq (1+\delta) \cdot \lambda_n \cdot J_{k^*}^2(q_{t,\epsilon}) \to 0 \quad (n \to \infty)$$

since $\lambda_n \to 0 \quad (n \to \infty)$.

In the eighth step of the proof we observe that the law of large numbers implies

$$T_{5,n} \to 0 \quad \text{in probability.}$$

In the ninth (and last) step of the proof we finish the proof by observing

$$\begin{aligned} T_{6,n} &= (1+2\delta+\delta^2)\cdot \mathbf{E}\left\{|q_{t,\epsilon}(X_t)-Y_t|^2\right\} - \mathbf{E}\left\{|q_t(X_t)-Y_t|^2\right\} \\ &= (2\delta+\delta^2)\cdot \mathbf{E}\left\{|q_{t,\epsilon}(X_t)-Y_t|^2\right\} + \int |q_{t,\epsilon}(x)-q_t(x)|^2 \mathbf{P}_{X_t}(dx) \\ &\leq (2\delta+\delta^2)\cdot (\max_{x\in\mathbb{R}^d} |q_{t,\epsilon}(x)|+L)^2+\epsilon. \end{aligned}$$

The right-hand side above can be made arbitrarily small by choice of δ and ϵ. □

3.4 Proof of Theorem 2

In the sequel we will show

$$\int |q_{n,s}(x)-q_s(x)|^2 \mathbf{P}_{X_s}(dx) = O_{\mathbf{P}}\left(C^{\frac{d}{2k^*+d}}\cdot\left(\frac{\log(n)}{n}\right)^{\frac{2k^*}{2k^*+d}}\right) \tag{16}$$

for all $s\in\{0,1,\dots,T\}$.

For $s=T$ we have $q_{n,T}(x)=0=q_T(x)$, so the assertion is trivial. So let $t<T$ and assume that the assertion holds for $s\in\{t+1,\dots,T\}$. By induction it suffices to show (16) for $s=t$, which we will show in the sequel in seven steps.

In the first step of the proof we show

$$\int |q_{n,t}(x)-q_t(x)|^2\mathbf{P}_{X_t}(dx) = O_{\mathbf{P}}\left(\frac{1}{n_t}\sum_{i=n_l+1}^{n}|q_{n,t}(X_{i,t}^{(t)})-q_t(X_{i,t}^{(t)})|^2+\frac{\log|\mathcal{P}_n|}{n_t}\right).$$

Let $\mathcal{D}_{n,t}$ be the set of all $X_{j,s}^{(r)}$ with either $r\geq t+1, s\in\{0,\dots,T\}$ and $j\in\{1,\dots,n\}$ or $r=t, s=t$ and $j\in\{1,\dots,n_l\}$. Conditioned on $\mathcal{D}_{n,t}$,

$$\{q_{n_l,t}^{(k,\lambda)} : (k,\lambda)\in\mathcal{P}_n\}$$

consists of $|\mathcal{P}_n|$ different functions. Furthermore, because of boundedness of $q_{n_l,t}^{(k,\lambda)}$ and q_t by L we have

$$\begin{aligned} \sigma^2_{(k,\lambda)} &:= \mathbf{Var}\{|q_{n_l,t}^{(k,\lambda)}(X_{n_l+1,t}^{(t)})-q_t(X_{n_l+1,t}^{(t)})|^2|\mathcal{D}_{n,t}\} \\ &\leq \mathbf{E}\{|q_{n_l,t}^{(k,\lambda)}(X_{n_l+1,t}^{(t)})-q_t(X_{n_l+1,t}^{(t)})|^4|\mathcal{D}_{n,t}\} \\ &\leq 4L^2 \int |q_{n_l,t}^{(k,\lambda)}(x)-q_t(x)|^2\,\mathbf{P}_{X_t}(dx). \end{aligned}$$

Using this and the Bernstein inequality (cf., e.g., Lemma A.2 in Györfi et al. (2002)) we get with the notation $\epsilon_n = c_1 \cdot \log |\mathcal{P}_n|/n_t$:

$$\mathbf{P}\{\int |q_{n,t}(x) - q_t(x)|^2 \, \mathbf{P}_{X_t}(dx) > (4L^2+1) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n,t}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 + \epsilon_n \, |\mathcal{D}_{n,t}\}$$

$$\leq |\mathcal{P}_n| \max_{(k,\lambda)\in\mathcal{P}_n} \mathbf{P}\{\int |q_{n_l,t}^{(k,\lambda)}(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx)$$
$$> (4L^2+1) \cdot \frac{1}{n} \sum_{i=n_l+1}^{n_t} |q_{n_l,t}^{(k,\lambda)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 + \epsilon_n \mid \mathcal{D}_{n,t}\}$$

$$\leq |\mathcal{P}_n| \cdot \max_{(k,\lambda)\in\mathcal{P}_n} \mathbf{P}\{\int |q_{n_l,t}^{(k,\lambda)}(x) - q_t(x)|^2 \, \mathbf{P}_{X_t}(dx) - \frac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n_l,t}^{(k,\lambda)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2$$
$$> \frac{1}{4L^2+1} \cdot \left(\sigma_{(k,\lambda)}^2 + \epsilon_n\right) |\mathcal{D}_{n,t}\}$$

$$\leq |\mathcal{P}_n| \cdot \max_{(k,\lambda)\in\mathcal{P}_n} \exp\left(\frac{-n_t \cdot \left(\frac{\sigma_{(k,\lambda)}^2 + \epsilon_n}{4L^2+1}\right)^2}{2\sigma_{(k,\lambda)}^2 + 2\frac{\sigma_{(k,\lambda)}^2+\epsilon_n}{4L^2+1} \cdot \frac{4L^2}{3}} \right)$$
$$\leq |\mathcal{P}_n| \cdot \max_{(k,\lambda)\in\mathcal{P}_n} \exp\left(- \frac{n_t(\sigma_{(k,\lambda)}^2 + \epsilon_n)}{2(4L^2+1)^2 + 2(4L^2+1) \cdot \frac{4L^2}{3}} \right)$$

$$\leq |\mathcal{P}_n| \cdot \exp\left(- \frac{c_1}{2(4L^2+1)^2 + 2(4L^2+1) \cdot \frac{4L^2}{3}} \cdot \log |\mathcal{P}_n| \right)$$

$$\to 0 \qquad (n \to \infty)$$

provided we choose c_1 sufficiently large.

In the second step of the proof we observe that as in the second step of the proof of Theorem 1 we have

$$\frac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n,t}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2$$
$$= O_{\mathbf{P}}\Bigg(\frac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n,t+1}(X_{i,t+1}^{(t)}) - q_{t+1}(X_{i,t+1}^{(t)})|^2 + \frac{\log |\mathcal{P}_n|}{n_t}$$
$$+ \min_{(k,\lambda)\in\mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n_l,t}^{(k,\lambda)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 \Bigg).$$

In the third step of the proof we show

$$\frac{1}{n_t} \sum_{i=n_l+1}^{n} |q_{n,t+1}(X_{i,t+1}^{(t)}) - q_{t+1}(X_{i,t+1}^{(t)})|^2 = O_{\mathbf{P}}\left(\int |q_{n,t+1}(x) - q_{t+1}(x)|^2 \mathbf{P}_{X_{t+1}}(dx) + \frac{\log |\mathcal{P}_n|}{n_t} \right).$$

Using

$$\mathbf{P}\{\frac{1}{n_t}\sum_{i=n_l+1}^{n}|q_{n,t+1}(X_{i,t+1}^{(t)})-q_{t+1}(X_{i,t+1}^{(t)})|^2$$
$$> (4L^2+1)\int|q_{n,t+1}(x)-q_{t+1}(x)|^2\mathbf{P}_{X_{t+1}}(dx)+\epsilon_n|\mathcal{D}_{n,t}\}$$
$$=\mathbf{P}\{\frac{1}{n_t}\sum_{i=n_l+1}^{n}|q_{n,t+1}(X_{i,t+1}^{(t)})-q_{t+1}(X_{i,t+1}^{(t)})|^2-\int|q_{n,t+1}(x)-q_{t+1}(x)^2\mathbf{P}_{X_{t+1}}(dx)$$
$$>4L^2\cdot\int|q_{n,t+1}(x)-q_{t+1}(x)|^2\mathbf{P}_{X_{t+1}}(dx)+\epsilon_n|\mathcal{D}_{n,t}\}$$

this follows as in the first step by an application of Bernstein inequality.

Let $k^*\in\{\lceil\frac{d}{2}\rceil,\ldots,K\}$ be as in Theorem 2 (i.e., $J_{k^*}^2(q_t)\le C$). Set

$$\bar{\lambda}_n=C^{-\frac{2k^*}{2k^*+d}}\cdot\left(\frac{\log n}{n}\right)^{\frac{2k^*}{2k^*+d}}$$

and choose (for n sufficiently large) λ_n^* such that

$$(k^*,\lambda_n^*)\in\mathcal{P}_n \text{ and } |\bar{\lambda}_n-\lambda_n^*|\le\frac{1}{n}.$$

In the fourth step of the proof we show

$$\min_{(k,\lambda)\in\mathcal{P}_n}\frac{1}{n_t}\sum_{i=n_l+1}^{n}|q_{n_l,t}^{(k,\lambda)}(X_{i,t}^{(t)})-q_t(X_{i,t}^{(t)})|^2=O_{\mathbf{P}}\left(\int|q_{n_l,t}^{(k^*,\lambda_n^*)}(x)-q_t(x)|^2\mathbf{P}_{X_t}(dx)+\frac{\log|\mathcal{P}_n|}{n_t}\right).$$

To see this, we observe that we have as in the third step of the proof

$$\frac{1}{n_t}\sum_{i=n_l+1}^{n}|q_{n_l,t}^{(k^*,\lambda_n^*)}(X_{i,t}^{(t)})-q_t(X_{i,t}^{(t)})|^2=O_{\mathbf{P}}\left(\int|q_{n_l,t}^{(k^*,\lambda_n^*)}(x)-q_t(x)|^2\mathbf{P}_{X_t}(dx)+\frac{\log|\mathcal{P}_n|}{n_t}\right),$$

hence the assertion follows from

$$\min_{(k,\lambda)\in\mathcal{P}_n}\frac{1}{n_t}\sum_{i=n_l+1}^{n}|q_{n_l,t}^{(k,\lambda)}(X_{i,t}^{(t)})-q_t(X_{i,t}^{(t)})|^2\le\frac{1}{n_t}\sum_{i=n_l+1}^{n}|q_{n_l,t}^{(k^*,\lambda_n^*)}(X_{i,t}^{(t)})-q_t(X_{i,t}^{(t)})|^2.$$

In the fifth step of the proof we show

$$\int|q_{n_l,t}^{(k^*,\lambda_n^*)}(x)-q_t(x)|^2\mathbf{P}_{X_t}(dx)$$
$$=O_{\mathbf{P}}\left(\frac{1}{n_l}\sum_{i=1}^{n_l}|q_{n_l,t}^{(k^*,\lambda_n^*)}(X_{i,t}^{(t)})-q_t(X_{i,t}^{(t)})|^2+\lambda_n^*\cdot J_{k^*}^2\left(\tilde{q}_{n_l,t}^{(k^*,\lambda^*)}\right)+C^{\frac{d}{2k^*+d}}\cdot\left(\frac{\log\ n}{n}\right)^{\frac{2k^*}{2k^*+d}}\right).$$

To do this we observe that for $t=c_2\cdot C^{\frac{d}{2k^*+d}}\cdot\left(\frac{\log n}{n}\right)^{\frac{2k^*}{2k^*+d}}$ we have:

$$\mathbf{P}\Big\{ \int |q_{n_l,t}^{(k^*,\lambda_n^*)}(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx)$$
$$> 2 \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} \left| q_{n_l,t}^{(k^*,\lambda_n^*)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)}) \right|^2 + \lambda_n^* \cdot J_{k^*}^2(\tilde{q}_{n_l,t}^{(k^*,\lambda_n^*)}) + t \Big\}$$
$$\le \sum_{j=0}^{\infty} \mathbf{P}\{2^j t \cdot I_{\{j \ne 0\}} \le \lambda_n^* \cdot J_{k^*}^2(\tilde{q}_{n_l,t}^{(k^*,\lambda_n^*)}) < 2^{j+1} \cdot t, \ldots\}$$
$$\le \sum_{j=0}^{\infty} \mathbf{P}\{\exists f = T_L g : g \in W^{k^*}([-A,A]^d),\ J_{k^*}^2(g) < 2^{j+1} \cdot t/\lambda_n^* \text{ and}$$
$$\int |f(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx) > 2 \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} |f(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2 + 2^j \cdot t\}$$
$$= \sum_{j=0}^{\infty} P_{j,n}.$$

Fix $j \in \mathbb{N}_0$. The last condition inside $P_{j,n}$ implies

$$\frac{|\int |f(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx) - \frac{1}{n_l} \sum_{i=1}^{n_l} |f(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)})|^2|}{2^j t + \int |f(x) - q_t(x)|^2 \mathbf{P}_{X_t}(dx)} > \frac{1}{2},$$

therefore we can apply Lemma 4 in order to bound $P_{j,n}$. In order to do this we will show that the assumptions of this lemma are satisfied. The first inequality there is implied by $t \ge c_3/n$. With the bound on the covering number of

$$\{T_L f : f \in W^{k^*}([-A,A]^d), J_{k^*}^2(f) \le 2^{j+1} \cdot t/\lambda_n^*\}$$

given in Lemma 3 we see that the second assumption of this lemma is implied by

$$c_4 \cdot \delta \cdot \sqrt{n} \ge \int_{c_5 \cdot \delta}^{\sqrt{\delta}} \left(\left(\left(\frac{\sqrt{2^{j+1} \cdot t/\lambda_n^*}}{u} \right)^{d/k^*} + 1 \right) \cdot \log \frac{64 e L^2 n}{(u/(4L^2))^2} \right)^{1/2} du \qquad (17)$$

for all $\delta \ge 2L^2 \cdot 2^j t$.

By choice of t we can bound the value of u inside the logarithm from below by $\frac{c_6}{n}$, therefore the right -hand side of (17) is bounded from above by

$$c_7 \cdot \left(\left(\frac{2^j \cdot t}{\lambda_n^*} \right)^{\frac{d}{4k^*}} \cdot (\sqrt{\delta})^{1 - \frac{d}{2k^*}} + 1 \right) \cdot \sqrt{\log(n)}.$$

So in order to show (17) it suffices to show

$$\delta^{\frac{1}{2} + \frac{d}{4k^*}} > c_8 \cdot \left(\frac{2^j \cdot t}{\lambda_n^*} \right)^{\frac{d}{4k^*}} \cdot \sqrt{\frac{\log(n)}{n}} \quad \text{and} \quad \delta \ge c_8 \sqrt{\frac{\log n}{n}}$$

for all $\delta \geq 2L^2 \cdot 2^j t$, which follows from $t = c_2 \cdot C^{d/(2k^*+d)} \cdot (\log n/n)^{2k^*/(2k^*+d)}$.

Hence by Lemma 4 we get

$$P_{j,n} \leq 15 \cdot \exp(-c_9 \cdot n_l \cdot 2^j \cdot t)$$

which implies

$$\sum_{j=0}^{\infty} P_{j,n} \leq 15 \cdot \frac{\exp(-c_9 \cdot n_l \cdot t)}{1 - \exp(-c_9 \cdot n_l \cdot t)} \to 0 \quad (n \to \infty).$$

In the sixth step of the proof we show

$$\frac{1}{n_l} \sum_{i=1}^{n_l} \left| q_{n_l,t}^{(k^*,\lambda_n^*)}(X_{i,t}^{(t)}) - q_t(X_{i,t}^{(t)}) \right|^2 + \lambda_n^* \cdot J_{k^*}^2 \left(\tilde{q}_{n_l,t}^{(k^*,\lambda_n^*)} \right)$$
$$= O_{\mathbf{P}} \left(\lambda_n^* \cdot J_{k^*}^2(q_t) + \frac{\log n_l}{n_l} \cdot (\lambda_n^*)^{-\frac{d}{2k^*}} + \frac{1}{n_l} \sum_{i=1}^{n_l} |q_{n,t+1}(X_{i,t+1}^{(t)}) - q_t(X_{i,t+1}^{(t)})|^2 + \frac{\log n_l}{n_l} \right)$$

This follows from Lemma 1 if we apply it (conditioned on all $X_{j,s}^{(r)}$ with $r \geq t+1$, $s \in \{0, \ldots, T\}$ and $j \in \{1, \ldots, n\}$) with

$$X_i = X_{i,t}^{(t)}, Y_i = \max\{f_{t+1}(X_{i,t+1}^{(t)}), q_{t+1}(X_{i,t+1}^{(t)})\} \text{ and } \bar{Y}_{i,n} = \max\{f_{t+1}(X_{i,t+1}^{(t)}), q_{n,t+1}(X_{i,t+1}^{(t)})\}$$

and if we observe

$$\frac{1}{n_l} \sum_{i=1}^{n_l} |Y_i - \bar{Y}_{i,n}|^2 \leq \frac{1}{n_l} \sum_{i=1}^{n_l} |q_{n,t+1}(X_{i,t+1}^{(t)}) - q_{t+1}(X_{i,t+1}^{(t)})|^2.$$

In the seventh step of the proof we observe

$$\frac{1}{n_l} \sum_{i=1}^{n_l} |q_{n,t+1}(X_{i,t+1}^{(t)}) - q_t(X_{i,t+1}^{(t)})|^2 = O_{\mathbf{P}}(\int |q_{n,t+1}(x) - q_{t+1}(x)|^2 \mathbf{P}_{X_{t+1}}(dx) + \frac{\log |\mathcal{P}_n|}{n_l}).$$

To see this, we condition on all data points $X_{j,s}^{(r)}$ with $r \geq t+1, s \in \{0, \ldots, T\}$ and $j \in \{1, \ldots, n\}$. Then the assertion follows by an application of Bernstein inequality as in steps 1 and 3.

The proof is complete. □

Acknowledgement

The author wishes to thank Nebojsa Todorovic for various helpful comments.

References

[1] Chow, Y. S. and Robbins, H. and Siegmund, D. (1971) *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, Boston.

[2] Cover, T. (1968). Rates of convergence for nearest neighbor procedures. In: *Proceedings of the Hawaii International Conference on System Sciences*, pp. 413-415, Honolulu, HI.

[3] Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **61**, pp. 467-481.

[4] Dippon, J. and Winter, S. (2006). Smoothing spline regression estimates for randomly right censored data. Preprint, University of Stuttgart.

[5] Duchon, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *R.A.I.R.O. Analyse Numérique* **10**, 5–12.

[6] Egloff, D. (2005) Monte Carlo Algorithms for Optimal Stopping and Statistical Learning. *Annals of Applied Probability* **15**, pp. 1-37.

[7] Egloff, D., Kohler, M., and Todorovic, N. (2006). A dynamic look-ahead Monte Carlo algorithm for pricing American options. Submitted for publication.

[8] Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

[9] Franke, J. and Diagne, M. (2002). Estimating market risk with neural networks. Technical report, University of Kaiserslautern. To appear in *Statistics & Decision* 2006/07.

[10] Glasserman, P. (2004). *Monte Carlo methods in financial engineering.* Springer.

[11] Glasserman, P., and Yu, B. (2004). Number of Paths versus Number of Basis Functions in American Option Pricing. *Annals of Applied Probability* **14**, pp. 1-30.

[12] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics, Springer.

[13] Kohler, M. (2006). Nonparametric regression with additional measurement errors in the dependent variable. *Journal of Statistical Planning and Inference* **136**, pp. 3339-3361.

[14] Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory* **47**, 3054–3058.

[15] Kohler, M., Krzyżak, A., and Schäfer, D. (2002). Application of structural risk minimization to multivariate smoothing spline regression estimates. *Bernoulli* **8**, pp. 475-489.

[16] Kohler, M, Kul, S., and Mathé, K. (2004). Least squares estimates for random design regression. Submitted for publication.

[17] Longstaff, F. A., and Schwartz, E. S. (2001). Valuing American Options by Simulation: A Simple Least-Square Approach. *Review of Financial Studies* **14**, pp. 113-147.

[18] Shiryayev, A. N. (1978). *Optimal Stopping Rules.* Applications of Mathematics, Springer Verlag.

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

[20] Tsitsiklis, J. N., and Van Roy, B. (1999). Optimal Stopping of Markov Processes: Hilbert Space Theory, Approximation Algorithms, and an Application to Pricing High-Dimensional Financial Derivatives. *IEEE Trans Autom. Control* **44**, pp. 1840-1851.

[21] van de Geer, S. (1987). A new approach to least squares estimation, with applications. *Ann. Statist.* **15**, 587–602.

[22] van de Geer, S. (1988). *Regression Analysis and Empirical Processes.* CWI-tract 45, Centre for Mathematics and Computer Science, Amsterdam.

[23] van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18**, 907–924.

[24] van de Geer, S. (2000). *Empirical Processes in M–estimation.* Cambridge University Press.

[25] Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.