# CONVERGENCE OF MONTE CARLO ALGORITHMS FOR PRICING AMERICAN OPTIONS

DANIEL EGLOFF AND MAUNG MIN-OO

Abstract. In this paper we study the convergence of the Longstaff-Schwartz algorithm for the valuation of American options. Our approach is based on empirical risk minimization initiated by Vapnik and Chervonenkis in the early 1970's and empirical processes techniques. This allows us to prove convergence, derive error estimates and a Central Limit Theorem for the sample estimators. It also opens up a variety of extensions and generalizations.

## 1. Introduction

### 1.1. **Valuation of American-style Options.**

Many financial products contain early exercise features which significantly contribute to their value. Therefore the valuation of American options, or more generally, of products with early exercise features has been considered as an important problem in Computational Finance.

Several methods have been developed to numerically solve the related optimal stopping problem. They range from binomial trees [9], Markov chain approximations [25], to semi-analytical approximations [2], direct integral equation methods, and PDE methods on the basis of variational inequalities, [3, 20], the linear complementary problem [19, 10], or the free boundary value problem [43].

An increase in product sophistication and model complexity during the last decade has called for new methods which can value American-style options with complicated exercise structures and depending on a multitude of risk factors. This stimulated the development of Monte Carlo methods to evade the curse of dimensionality caused by these new products and models. Various approaches have been proposed. The first landmark papers in this direction are [6, 5, 40, 7]. The state of development as of 1998 is described in the overview paper [8].

Recently, in 1999, Longstaff and Scwartz introduced a new Monte Carlo approach [28]. Their method is based on a parametric approximation scheme for Bermudan options in discrete time. They showed how to calculate the parameters algorithmically by solving a sequence of least square problems. A brief sketch of a proof for convergence of the algorithm is then outlined. A more detailed analysis of the convergence proof and a Central Limit Theorem is then discussed in the work of Clément, Lamberton and Protter [37]. However, their approach is different from the methods used in this paper.

It is also worth noting that Tsitsiklis and Van Roy [41] independently proposed a simpler algorithm for infinite horizon discrete time optimal stopping problem using also parametric approximations. Their approach is based on stochastic approximation techniques [23, 4, 24] and the parameters are calculated by temporal difference updates.

*Date*: 21st August 2002.

2000 *Mathematics Subject Classification.* Primary 91B28, 60G40, 93E20; Secondary 65C05, 93E24, 62G05.

*Key words and phrases.* Optimal stopping, American options, empirical processes, uniform law of large numbers, Vapnik-Chervonenkis classes, Monte Carlo methods.

The second author is partially supported by a research grant from NSERC.

All of the Monte Carlo algorithms we referred to so far, approximate the value function or the early exercise rule in some way, hence provide a lower bound to the true option value. In contrast to this, a recent paper of Rogers [17] focuses on the dual problem to calculate upper bounds. Finally, a recent comparative study of various Monte Carlo approaches can be found in [27].

1.2. **Our Approach and Contributions.** In this paper, we analyze the convergence and error estimates of the Longstaff-Schwartz algorithm, which is based on a linear approximation of the so called $q$-function, which plays a key role for optimal stopping problems and can be thought of as the live value of the option if not exercised immediately. The overall approximation error of the algorithm is decomposed into two main errors: the approximation error, caused by the finite resolution of the approximation architecture, and a stochastic error, or the sample error, which is caused by the finite sample size of the Monte Carlo method.

The approximation error is of a deterministic nature and its convergence is controlled by functional analytic properties of the approximating architecture and and the degree of smoothness of the $q$-function. Results of G. Freud and H. N. Mhaskar [30] on weighted polynomial approximation can be applied to prove convergence without restricting ourselves to compact domains of the state space.

The sample error, which is our main concern in this paper, is analyzed in the framework of *empirical risk minimization*, which has been promoted by Vapnik and Chervonenkis in a series of papers [44, 45, 46] since the early 1970's.

Let $X$ be a random variable defined on a probability space $(\Omega, P)$, taking values in an arbitrary set $S$, and let $\mathcal{G}$ be a class of functions defined on $S$. Consider a risk functional $L(g) = Pl(g(X)) \equiv E_P[l(g(X)]$ to be minimized. Empirical risk minimization approximates a minimizer $g^* = \arg\min_{g\in\mathcal{G}} L(g)$ by the sample minimizer $\hat{g}_n = \arg\min_{g\in\mathcal{G}} L_n(g)$ of the empirical criterion $L_n(g) = n^{-1}\sum_{i=1}^n l(g(X_i))$, based on $n$ samples from $X$. Empirical risk minimization is based on the inequalities

$$\begin{aligned} 0 &\le L(\hat{g}_n) - L(g^*) \le 2\sup_{g\in\mathcal{G}} |L_n(g) - L(g)|, \\ 0 &\le L_n(\hat{g}_n) - L(\hat{g}_n) \le \sup_{g\in\mathcal{G}} |L_n(g) - L(g)|. \end{aligned} \tag{1.1}$$

The first series of inequalities provide an upper bound on the sub-optimality of $\hat{g}_n$ for the risk function $L(g)$ over $\mathcal{G}$. The second give an upper bound on the error if the empirical criterion $L_n(\hat{g}_n)$ is used to estimate the risk $L(\hat{g}_n)$ of the sample minimizer $\hat{q}_n$.

If a Uniform Law of Large Numbers can be proved for the probability distribution $P$ and the function classes $\mathcal{G}$, such that the right hand sides of (1.1) converge to zero almost surely, we can conclude that $L_n(\hat{g}_n)$ converges to $L(g^*)$. Finally, convergence of the sample minimizes to the minimizer $g^*$ follows if the risk function satisfies an estimate

$$d(g, g^*) \le L(g) - L(g^*) \tag{1.2}$$

for some metric $d$ on $\mathcal{G}$. Note that a uniform one-sided estimate for $\sup_{g\in\mathcal{G}}\big(L_n(g) - L(g)\big)$, together with a point estimate for $L(\hat{g}_n) - L_n(\hat{g}_n)$ is sufficient to carry through the above arguments. This can result in tighter estimates as been emphasized by Vapnik [46].

The difficulty in proving convergence for the Longstaff-Schwartz algorithm is twofold. First, the sequence of sample minimizers is nested because of a backward recursion. The error of the previous step affects the error of the next step. The error propogation can be controlled provided some minor continuity assumptions are satisfied. The second source of difficulties is the expression of the risk function, which is quadratic but contains explicitly the approximated optimal stopping time. The simplified $Q$-value algorithm of Tsitsiklis and Van Roy avoids this difficulty by proposing simpler risk functions. We refer to section 3 for more details. Besides proving overall convergence of the Longstaff-Schwartz algorithm we derive detailed error estimates.

The article is organized as follows. In Section 2 we briefly review the no-arbitrage valuation of an American options in discrete time, i.e a Bermudan option, by solving an optimal stopping problem, introduce the notation and some basic path functionals that will be used. We also derive, in this section, a relatively simple estimate of the error caused by truncating the pay-off function from above.

After describing the Longstaff-Schwartz algorithm in the next Section 3, we analyze the approximation error in Section 4. The main estimates are obtained in Corollaries 4.5 and 4.6. After giving a brief review of the relevant concepts and results that we use from the theory of empirical processes in Section 5, we turn our attention in Section 6 to estimation of the sample error and complexity, which is the main concern of this paper. Convergence is proved in Theorem 6.8 and the main sample error bound is derived in Theorem 6.15.

The proof of the bound on the sample error utilizes Talagrand's uniform deviation inequality from the mean. It exhibits the unpleasant feature that the exponential rate of convergence deteriorates exponentially fast with increasing number of time steps. This however is a consequence of the lack of smoothness of the functional $y_{h,t}$ appearing in the loss function, which results in relatively poor point estimates (2.23). Our numerical simulations seem to indicate that the deterioration of the convergence rate with the number of time steps is inherent in these type of algorithms.

Since our estimates are based on a well-established general theory of empirical processes and risk minimization procedures our results and our methods actually apply to a wide range of pay-off functions and approximation schemes, although for the purpose of clarity, we restrict ourselves mainly to the Longstaff-Schwartz algorithm in this paper. A further merit of applying empirical process theory is a rather straightforward proof of a Central Limit Theorem for the sample minimizers, which we establish in Section 7, under the assumption of a bounded payoff function and $L_\infty$ approximation architecture. Our setup also makes a connection to the work of Arcones [1], who investigated limit behavior of approximate $M$-estimators.

Another advantage of the machinery of empirical risk minimization and empirical process theory is that it provides a coherent framework for analyzing more general Monte Carlo algorithms for optimal stopping problems. This would encompass more general approximation architectures such as radial basis functions, neural networks or $n$-terms approximants, also known as basis projection pursuit, all of which are non-linear, but also nonparametric approximation schemes such as kernel smoothing estimators or local polynomial regression.

*Structural risk minimization*, a generalization of empirical risk minimization, provides yet another dimension for generalization and bridges to the field of *model selection.* Structural risk minimization can be applied to answer the question of how to balance the different error contributions, to select the proper sieve and truncation level to minimize the overall error of the algorithm, given a fixed sample size. As a first result in the direction of structural risk minimization, a generalization of empirical risk minimization, we extend in Section 8 the Longstaff-Schwartz algorithm to sieve estimation and establish a convergence and consistency result. The final Section then discusses open problems and future directions for research.

1.3. **Important Remark on Notation.** In many cases we need to be precise with respect to which measure the integrations take place. We therefore apply the shorthand notation $PX = E_P[X]$ to denote the expectation of $X$ with respect to a measure $P$, as it is customary in empirical process theory, and analogously $P(X \mid \mathcal{F}) = E_P[X \mid \mathcal{F}]$ for the conditional expectation.

## 2. Review of American Options Pricing and Optimal Stopping

2.1. **American Options Pricing and Optimal Stopping Problem.** It is well known that the no-arbitrage price $v_0$ of an American option with payoff at exercise $f_t$ and expiry date $T$ can be calculated by solving the optimal stopping problem

$$v_0 = \sup_{\tau \in \mathcal{T}_{[0,T]}} P\Big(e^{-\int_0^\tau r(s)ds}\, f_\tau\Big), \tag{2.1}$$

where $r(t)$ is the spot rate process, $\mathcal{T}_{[0,T]}$ are the stopping times with values in $[0,T]$ and the expectation is taken with respect to the martingale measure $P$. A sufficient condition such that the expectation in (2.1) makes sense for all stopping times is that the discounted payoff process $\exp(-\int_0^t r(s)ds)\, f_t$ is of class $D$. For additional details we refer to [22, 34] and [21, Appendix D].

2.2. **Bermudan Approximations.** The first approximation is to replace the continuous time problem (2.1) by a discrete time approximation. Restricting exercise and trading dates to discretization dates

$$0 = t_0 < t_1 < \ldots < t_N = T \tag{2.2}$$

leads to a Bermudan approximation of the original American option. The finer the time discretization, the better the original American option is approximated. However, convergence of discrete time approximations to the continuous time limit is rather subtle and requires additional conditions, primarily on the filtration. We refer the interested reader to [26, 33].

2.3. **Discrete Time Optimal Stopping Problems.** Henceforth we fixed a suitable time discretization (2.2) and consider the related discrete time problem. Let $f_t$, $t = 0, \ldots, T$ be a discrete time payoff process, adapted to the filtration $\mathcal{F}_t$. Denote by $d_{t,s}$ the discount factor from $s$ back to $t$, which we assume to be $\mathcal{F}_t$-measurable. The option value $V_t$ at time $t$ allows the optimal stopping characterization

$$V_t = ess\sup\nolimits_{\tau \in \mathcal{T}_{t,\ldots,T}} P\big(d_{t,\tau}\, f_\tau \mid \mathcal{F}_t\big) = P\big(d_{t,\tau_t^*}\, f_{\tau_t^*} \mid \mathcal{F}_t\big), \tag{2.3}$$

where the optimal stopping times $\tau_t^*$ are given by

$$\tau_t^* = \inf\{s \geq t \mid V_s \leq f_s\}. \tag{2.4}$$

Closely related to $V_t$ is the quantity

$$Q_t = ess\sup\nolimits_{\tau \in \mathcal{T}_{t+1,\ldots,T}} P\big(d_{t,\tau}\, f_\tau \mid \mathcal{F}_t\big) = P\big(d_{t,\tau_{t+1}^*}\, f_{\tau_{t+1}^*} \mid \mathcal{F}_t\big), \tag{2.5}$$

called the $Q$-value, representing the optimal stopping value at time $t$, subject to the condition of not stopping immediately. It is easily verified by induction that

$$Q_t = P\big(d_{t,t+1} V_{t+1} \mid \mathcal{F}_t\big), \tag{2.6}$$

and

$$V_t = \max(f_t, Q_t). \tag{2.7}$$

From the last equation one sees that $Q_t$ can be interpreted as the continuation value of the option contract. For the sake of completeness, extend the definition of $Q_t$ by putting $Q_T \equiv f_T$.

2.4. **Dynamic Programming Principle.** The dynamic programming principle can be applied to calculate option value process recursively according to

$$\begin{aligned} V_T &= f_T, \\ V_t &= \max\big(f_t, P(d_{t,t+1} V_{t+1} | \mathcal{F}_t)\big), \quad t = T-1, \ldots, 0. \end{aligned} \tag{2.8}$$

Similarly, the $Q$-value $Q_t$ satisfies the backward recursion

$$\begin{aligned} Q_T &= f_T, \\ Q_t &= P\big(d_{t,t+1} \max(f_{t+1}, Q_{t+1}) | \mathcal{F}_t\big), \quad t = T-1, \ldots, 0. \end{aligned} \tag{2.9}$$

As already noted in [41], the expectation operator enters linearly in the backward recursion (2.9), and a sample based approximation thereof does not introduce a bias, hence promotes itself naturally for a sample approximation.

The optimal stopping times $\tau_0^* \leq \ldots \leq \tau_T^*$ obviously determine $V_t$ or $Q_t$. Conversely they can be recovered from the knowledge of either $V_t$ or $Q_t$ by the following backward recursion

$$\begin{aligned} \tau_T^* &= T, \\ \tau_t^* &= t\, 1_{\{V_t = f_t\}} + \tau_{t+1}^* 1_{\{V_t > f_t\}} = \\ &= t\, 1_{\{Q_t \leq f_t\}} + \tau_{t+1}^* 1_{\{Q_t > f_t\}}. \end{aligned} \tag{2.10}$$

Note that the last equality in (2.10) follows directly from (2.8) and (2.6).

2.5. **Payoff Truncation.** Option payoff functions in Finance are typically unbounded. On the other hand any numerical implementation works at finite precision. This calls for a truncation of the payoff function. For any cutoff level $c > 0$ introduce the truncation operator

$$\mathcal{T}_c : f \mapsto f 1_{\{|f| \leq c\}}. \tag{2.11}$$

Then, for $f \in L_p(\Omega, P)$, it follows that

$$\lim_{c \to \infty} c^p P\big(|f| > c\big) = 0. \tag{2.12}$$

The next results bounds the error of truncating the payoff function.

**Proposition 2.1.** *Assume that $f_t \in L_p(\Omega, \mathcal{F}_t, P)$ for all $t$. Let $Q(\mathcal{T}_c f)$ be the $Q$-value of the truncated payoff process $\mathcal{T}_c f_t$. Then for $t \leq T-1$,*

$$\|Q_t(f) - Q_t(\mathcal{T}_c f)\|_{p,P} \leq \sum_{s=t+1}^{T} \|f_s\|_{p,P}\, P(f_s > c)^{\frac{p-1}{p}} \leq o(c^{p-1}). \tag{2.13}$$

*Proof.* Apply the relation (2.9) and note that $|\max(a,x) - \max(a,y)| \leq |x - y|$. Then

$$\begin{aligned} &\|Q_t(f) - Q_t(\mathcal{T}_c f)\|_{p,P} \leq \\ &\qquad \|P\big(\max(f_{t+1}, Q_{t+1}(f)) - \max(\mathcal{T}_c f_{t+1}, Q_{t+1}(\mathcal{T}_c f)) \mid \mathcal{F}_t\big)\|_{p,P} \leq \\ &\quad \|P\big(f_{t+1} - \mathcal{T}_c f(X)_{t+1} \mid \mathcal{F}_t\big)\|_{p,P} + \|P\big(Q_{t+1}(f) - Q_{t+1}(\mathcal{T}_c f) \mid \mathcal{F}_t\big)\|_{p,P} \leq \\ &\qquad\qquad \|f_{t+1} 1_{\{f_{t+1} > c\}}\|_{p,P} + \|Q_{t+1}(f) - Q_{t+1}(\mathcal{T}_c f)\|_{p,P}. \end{aligned}$$

Apply Hölder's inequality on the first term and proceed by induction. □

2.6. **Markovian State Process.** It is customary to assume that the the payoff process is a function $f_t = f_t(X_t)$ of a discrete time Markov process $X_t$ with values in a potentially high-dimensional euclidian space $\mathbb{R}^d$ and that $\mathcal{F}_t = \sigma(X_t)$ is the natural $P$-completed filtration. This is no real restriction because path dependency can always be accounted for by adding additional state variables performing historical bookkeeping. The Markov property can then be exploited to represent the option value and the $Q$-value in terms of Borel measurable functions $v_t, q_t : \mathbb{R}^d \to \mathbb{R}$ such that

$$V_t = v_t(X_t), \quad Q_t = q_t(X_t).$$

The recursion (2.9) can be expressed as

$$q_t = P_{t,t+1}\big(d_{t,t+1} \max(f_{t+1}, q_{t+1})\big), \tag{2.14}$$

where $P_{t,t+1}$ is the transition function of the Markov process $X_t$. In analogy with the $Q$-value the $q_t$ are called $q$-functions.

2.7. **Basic Assumptions and Notations.** Assume that $X_t$, $t = 0, \dots, T$ is a discrete time Markov process with values in $\mathbb{R}^d$, defined on a probability space $(\Omega, P, \mathcal{F})$. Let $\mathcal{F}_t$ be the natural filtration of $X_t$ and

$$\mathcal{R} = \mathcal{R}_T = (\mathbb{R}^d)^{T+1}. \tag{2.15}$$

the path space of $X = (X_0, \dots, X_T)$. Denote by $\mu_{X_t}$ the law of $X_t$ on $\mathbb{R}^d$ and by $L_2(\mathbb{R}^d, \mu_{X_t})$ the space of $\mu_{X_t}$-square integrable Borel functions. Note that via the embedding $h \mapsto h \circ X_t$ we can always regard $L_2(\mathbb{R}^d, \mu_{X_t})$ as a subspace of $L_2(\Omega, \mathcal{F}, P)$.

Markov path functions are Borel measurable functions $h : \mathcal{R} \to \mathbb{R}^{T+1}$ defined on the path space $\mathcal{R}$ such that

$$h(x) = (h_0(x_0), \dots, h_T(x_T)), \tag{2.16}$$

for some real-valued Borel functions $h_t$ on $\mathbb{R}^d$. In the following we use $h(x)_t$ as a shorthand notation for $h_t(x_t)$ and identify a sequence of Borel functions $(h_0, \dots, h_T)$ with the corresponding Markov path function. In particular we view the payoff function and the $q$-function as Markov path function.

For $1 \le p \le \infty$ let $L_p(X)$ be the space of Markov path functions with $h_t \in L_p(\mathbb{R}^d, \mu_{X_t})$ for every $t = 0, \dots, T$, endowed with the norm

$$\|h\|_{p,X} = \|h\|_{1,p,X} = \sum_{t=0}^{T} \|h_t\|_{p,\mu_{X_t}}. \tag{2.17}$$

**Assumption 2.2.** *To account for any integrability issues assume from now on that the payoff process $f$ is nonnegative and that $f \in L_2(X)$.*

With this assumption, $Q_t$ and $V_t$ are square integrable and $q \in L_2(X)$.

2.8. **The Functionals $\tau_{h,t}$ and $y_{h,t}$.** For any Markov path function $h$ introduce the functional $\tau_{h,t}$ on $\mathcal{R}$ with values in $t, \dots, T$ defined as

$$\begin{aligned} \tau_{h,t}(x) &= \inf\{s \ge t \mid h(x)_s \le f(x)_s\} \wedge T \\ &= t\, 1_{\{h(x)_t \le f(x)_t\}} + \tau_{h,t+1}(x)\, 1_{\{h(x)_t > f(x)_t\}}, \end{aligned} \tag{2.18}$$

and the functional $y_{h,t}$

$$y_{h,t}(x) = d_{t,\tau_{h,t}(x)} f(x)_{\tau_{h,t}(x)} = d_{t,\tau_{h,t}(x)} f_{\tau_{h,t}(x)}(x_{\tau_{h,t}(x)}). \tag{2.19}$$

Note that $\tau_{h,t}$ and $y_{h,t}$ only depend on $x_s$ and $h_s$ for $s \ge t$, and that $\tau_{h,T}(x) = T$. It follows that

$$q(X)_t = P\big(y_{q,t+1}(X) \mid \mathcal{F}_t\big) \tag{2.20}$$

and that the optimal stopping times $\tau_t^*$ are given by $\tau_t^* = \tau_{q,t}(X)$.

It is fundamental to optimal stopping that $\tau_{h,t}$ is not continuous on all of $\mathcal{R}$. To see this let $x \in \mathcal{R}$ satisfy $h(x)_u = f(x)_u$ and $h(x+y)_u > f(x+y)_u$ for $\|y\|$ small enough as well as $h(x)_v = f(x)_v$ for $t < u < v$. Then if $\tau_{h,t}$ would be continuous at $x$, $\tau_{h,t}$ must be constant on a sufficiently small neighborhood of $x$. But by construction, this is not the case because $\tau_{h,t}(x)=u$ whereas $\tau_{h,t}(x+y)=v$ for any $y$ arbitrarily small.

Our error analysis depends heavily on the properties of the functionals $\tau_{h,t}$ and $y_{h,t}$. The first important observation is that we can control $L_2$-norm of the projection of $y_{h,t+1}(X)-y_{q,t+1}(X)$ onto $L_2(\Omega,\mathcal{F}_t,P)$.

**Proposition 2.3.** *Let $q$ denote the q-function and assume $h\in L_2(X)$. Then*

$$\begin{aligned}\|P\big(y_{h,t+1}(X)-y_{q,t+1}(X)\mid\mathcal{F}_t\big)\|_{2,P}\le&\\ \|h(X)_{t+1}-q(X)_{t+1}\|_{2,P}+\|P\big(y_{h,t+2}(X)-y_{q,t+2}(X)\mid\mathcal{F}_{t+1}\big)\|_{2,P}&\qquad(2.21)\end{aligned}$$

*Proof.* We can assume that $d_{t,s}=1$, which will simplify notations.

$$\begin{aligned}&\|P\big(y_{h,t+1}(X)-y_{q,t+1}(X)\mid\mathcal{F}_t\big)\|_{2,P}\le\\ &\quad\|P\big((f(X)_{t+1}-Q_{t+1})(1_{\{q(X)_{t+1}\le f(X)_{t+1}\}}-1_{\{h(X)_{t+1}\le f(X)_{t+1}\}})\mid\mathcal{F}_t\big)\|_{2,P}+\\ &\quad\|P\big(f(X)_{\tau_{q,t+2}(X)}1_{\{q(X)_{t+1}>f(X)_{t+1}\}}-f(X)_{\tau_{h,t+2}(X)}1_{\{h(X)_{t+1}>f(X)_{t+1}\}}+\\ &\qquad Q_{t+1}(1_{\{q(X)_{t+1}\le f(X)_{t+1}\}}-1_{\{h(X)_{t+1}\le f(X)_{t+1}\}})\mid\mathcal{F}_t\big)\|_{2,P}=I_1+I_2.\end{aligned}$$

The first term can be estimated further as follows

$$\begin{aligned}I_1=&\|P\big((f(X)_{t+1}-Q_{t+1})(1_{\{q(X)_{t+1}\le f(X)_{t+1}<h(X)_{t+1}\}}-\\ &\qquad1_{\{h(X)_{t+1}\le f(X)_{t+1}<q(X)_{t+1}\}})\mid\mathcal{F}_t\big)\|_{2,P}\le\\ &\|P\big((f(X)_{t+1}-Q_{t+1})(1_{\{0\le f(X)_{t+1}-q(X)_{t+1}<h(X)_{t+1}-q(X)_{t+1}\}}-\\ &\qquad1_{\{h(X)_{t+1}-q(X)_{t+1}\le f(X)_{t+1}-q(X)_{t+1}<0\}})\mid\mathcal{F}_t\big)\|_{2,P}\le\\ &\qquad\|P\big(|h(X)_{t+1}-q(X)_{t+1}|\mid\mathcal{F}_t\big)\|_{2,P}\le\|h(X)_{t+1}-q(X)_{t+1}\|_{2,P}.\end{aligned}$$

To assess the second term remember the definition (2.5) of $Q_{t+1}$, which is in terms of $\tau_{q,t+2}$ can be written as

$$Q_{t+1}=P\big(d_{t+1,\tau_{q,t+2}(X)}\,f(X)_{\tau_{q,t+2}(X)}\mid\mathcal{F}_t\big).$$

Therefore

$$\begin{aligned}I_2=&\|P\big(f(X)_{\tau_{q,t+2}(X)}(1-1_{\{h(X)_{t+1}\le f(X)_{t+1}\}})-\\ &\qquad f(X)_{\tau_{h,t+2}(X)}1_{\{h(X)_{t+1}>f(X)_{t+1}\}}\mid\mathcal{F}_t\big)\|_{2,P}\le\\ &\|P\big((f(X)_{\tau_{q,t+2}(X)}-f(X)_{\tau_{h,t+2}(X)})1_{\{h(X)_{t+1}>f(X)_{t+1}\}}\mid\mathcal{F}_t\big)\|_{2,P}\le\\ &\qquad\qquad\|P\big(y_{h,t+2}(X)-y_{q,t+2}(X)\mid\mathcal{F}_{t+1}\big)\|_{2,P}.\end{aligned}$$

□

Proposition 2.3 only controls the deviation of $y_{h,t}$ from $y_{q,t}$. The next proposition provides more general point-wise estimates for the functionals $\tau_{h,t}$ and $y_{h,t}$.

**Proposition 2.4.** *Let $g,h:\mathcal{R}\to\mathbb{R}$. Then*

$$|\tau_{g,t}(x)-\tau_{h,t}(x)|\le\sum_{s=t}^{T-1}\big(s+\tau_{h,s+1}(x)\big)1_{\{|f(x)_s-h(x)_s|\le|g(x)_s-h(x)_s|\}}\qquad(2.22)$$

$$|y_{g,t}(x)-y_{h,t}(x)|\le\sum_{s=t}^{T-1}\Big(\sum_{r=s}^{T}|f(x)_r|\Big)1_{\{|f(x)_s-h(x)_s|\le|g(x)_s-h(x)_s|\}}\qquad(2.23)$$

*Proof.* The definition of $\tau_{h,t}$ shows that

$$\begin{aligned}|\tau_{g,t}(x)-\tau_{h,t}(x)|\le&\,t|1_{\{g(x)_t\le f(x)_t\}}-1_{\{h(x)_t\le f(x)_t\}}|+\\ &|\tau_{g,t+1}(x)1_{\{g(x)_t>f(x)_t\}}-\tau_{h,t+1}(x)1_{\{h(x)_t>f(x)_t\}}|=t\,I_1+I_2.\end{aligned}$$

$$I_1 = |1_{\{g(x)_t \le f(x)_t < h(x)_t\}} - 1_{\{h(x)_t \le f(x)_t < g(x)_t\}}| \; = \\ |1_{\{g(x)_t - h(x)_t \le f(x)_t - h(x)_t < 0\}} - 1_{\{0 \le f(x)_t - h(x)_t < g(x)_t - h(x)_t\}}| \; \le \\ 1_{\{|f(x)_t - h(x)_t| \le |g(x)_t - h(x)_t|\}}.$$

As for $I_2$,

$$I_2 \le |\big(\tau_{g,t+1}(x) - \tau_{h,t+1}(x)\big) 1_{\{g(x)_t \le f(x)_t\}}| + \\ |\tau_{h,t+1}(x)\big(1_{\{g(x)_t > f(x)_t\}} - 1_{\{h(x)_t > f(x)_t\}}\big)| \; \le \\ |\tau_{g,t+1}(x) - \tau_{h,t+1}(x)| + \tau_{h,t+1}(x)|1_{\{h(x)_t \le f(x)_t\}} - 1_{\{g(x)_t \le f(x)_t\}}|,$$

where $1_{\{g(x)_t > f(x)_t\}}$ has been replaced by $1 - 1_{\{g(x)_t \le f(x)_t\}}$ and similar for $h$. The last term on the right hand side can then be estimated as for $I_1$. By induction, this shows (2.22). As for (2.23), along the same lines one obtains

$$|y_{g,t}(x) - y_{h,t}(x)| \le \\ \big(|f(x)_t| + |y_{h,t+1}(x)|\big) 1_{\{|f(x)_t - h(x)_t| \le |g(x)_t - h(x)_t|\}} + |y_{g,t+1}(x) - y_{h,t+1}(x)|.$$

Now note that $|y_{h,t+1}(x)| \le \sum_{s=t+1}^{T-1} |f(x)_s|$ and complete by induction. □

## 3. The Algorithm

The previous discussion shows that the $q$ function is obtained by applying the conditional expectation operator $P(. \mid \sigma(X_t))$ to either $d_{t,\tau^*_{t+1}} f_{\tau^*_{t+1}}$ or alternatively to $\max(f_{t+1}, q_{t+1})(X_{t+1})$. Now on the Hilbert spaces $L_2(\Omega, \mathcal{F}, P)$

$$P(. \mid \sigma(X_t)) : L_2(\Omega, \mathcal{F}, P) \to L_2(\Omega, \sigma(X_t), P) \tag{3.1}$$

are orthogonal projection operator. The key idea of Longstaff-Schwartz algorithm and the related algorithms $Q$-value algorithm is to compose this orthogonal projection operator with a projection onto suitable finite dimensional subspaces. These finite dimensional projections are then further approximated by Monte Carlo methods.

3.1. **Approximation Architectures.** An approximation architecture is represented by a sequence of subspaces

$$\mathcal{H}_t \subset L_2(\mathbb{R}^d, \mu_{X_t}). \tag{3.2}$$

The corresponding subspace of square integrable Markov path function is denoted by $\mathcal{H} \subset L_2(X)$. The approximation architecture is called linear if $\mathcal{H}_t$ are linear subspaces, and write $\dim(\mathcal{H}) = k$ if $\dim(\mathcal{H}_t) = k$ for all $t$. Linear approximation architectures can be constructed for example by means of ordinary multivariate polynomials, Legendre, Hermite or Laguerre polynomials, or more generally, $\mathcal{H}_t$ can be chosen as the span of the first $k$ elements of a Riesz basis[1] for $L_2(\mathbb{R}^d, \mu_{X_t})$.

3.2. **The Algorithm of Longstaff and Schwartz.** Let $\mathcal{H}$ be a linear approximation architecture. Define $q_{\mathcal{H}} \in \mathcal{H}$ recursively by

$$q_{\mathcal{H}}(X)_t = \mathrm{pr}_{\mathcal{H}_t}\big(y_{q_{\mathcal{H}},t+1}(X)\big), \tag{3.3}$$

where $y_{q_{\mathcal{H}},t+1}(X) \in L_2(\Omega, \mathcal{F}, P)$ is defined in (2.19) and $\mathrm{pr}_{\mathcal{H}_t}$ is the orthogonal projection onto $\mathcal{H}_t \subset L_2(\mathbb{R}^d, \mu_{X_t}) \hookrightarrow L_2(\Omega, \mathcal{F}, P)$. Definition (3.3) makes sense because $y_{q_{\mathcal{H}},t+1}(X)$ only depends on $q_{\mathcal{H}}(X)_s$ and $X_s$ for $s \ge t+1$. Let

$$q_{\mathcal{H},t} = \underset{h_t \in \mathcal{H}_t}{\arg\min}\, P\big(h_t(X_t) - y_{q_{\mathcal{H}},t+1}(X)\big)^2 \tag{3.4}$$

[1] A Riesz basis of a Hilbert space $H$ is a system $\{x_n\} \subset H$ such that there exists an orthonormal basis $\{e_n\} \subset H$ and a bounded and boundedly invertible map $T$ of $H$ with $Te_n = x_n$ for all $n$.

be the variational characterization of $q_{\mathcal{H}}$. A sample approximation is now obtained by replacing the error functional in (3.4) by its empirical counterpart such that

$$\hat{q}_{n,\mathcal{H},t} = \underset{h_t \in \mathcal{H}_t}{\arg\min}\, P_n \big(h_t(X_t) - y_{\hat{q}_{n,\mathcal{H}},t+1}(X)\big)^2, \tag{3.5}$$

where $P_n$ be the empirical measure based on $n$ independent sample paths of $X$. Expressing $\hat{q}_{n,\mathcal{H},t}$ in terms of a basis for $\mathcal{H}_t$, the minimization problem (3.5) is transformed into a least square regression problem. This parametrization has no impact on the convergence as it affects the estimates only by a constant. It is now apparent that the key step in the convergence proof is to understand the effect of the functional $y_{h,t}$.

3.3. **The $Q$-Value Algorithm.** The $Q$-value algorithm of Tsitsiklis and Van Roy [41] is a simplified version of the Longstaff-Schwarz algorithm. The projection operator is applied to $d_{t,t+1} \max(f_{t+1}, q_{t+1}(X_{t+1}))$, which is justified by (2.9). The recursion is then given by

$$\hat{q}_{n,\mathcal{H},t} = \underset{h_t \in \mathcal{H}_t}{\arg\min}\, P_n \big(h_t(X_t) - \max(f_{t+1}, \hat{q}_{n,\mathcal{H}}(X)_{t+1}\big)^2, \tag{3.6}$$

and the approximated option price is

$$\hat{v}_{\mathcal{H},0} = \max\big(f(X)_0, P(d_{0,1} \max(f(X)_1, \hat{q}_{n,\mathcal{H}}(X)_1)\big).$$

Note that there is no need in keeping track of the approximating optimal stopping times as in the case of the Longstaff-Schwartz algorithm. The convergence proof of the $Q$-value algorithm is much simpler. Therefore, we will mainly focus on the analysis of the Longstaff-Schwartz algorithm.

3.4. **Error Decomposition.** The the overall error of the above algorithms can be decomposed into two components, the approximation error, a deterministic quantity caused by the finite resolution of the approximation architecture, and the sample error, which is stochastic and results from a finite sample approximation. More precisely,

$$\|q - \hat{q}_{n,\mathcal{H}}\|_{2,X} \le \|q - q_{\mathcal{H}}\|_{2,X} + \|q_{\mathcal{H}} - \hat{q}_{n,\mathcal{H}}\|_{2,X}, \tag{3.7}$$

of which the first term on the right hand side is the approximation error, and the second term the sample error. If the payoff is truncated according to proposition 2.1, a third error term appears.

## 4. Approximation Error

### 4.1. General Linear Approximation Architectures.

**Theorem 4.1.** *Let $\mathcal{H}$ be a linear approximation architecture such that the maximal error in approximating the q-function satisfies*

$$\|q(X)_t - \mathrm{pr}_{\mathcal{H}_t}\big(q(X)_t\big)\|_{2,P} \le \varepsilon_t. \tag{4.1}$$

*Then*

$$\|q(X)_t - q_{\mathcal{H}}(X)_t\|_{2,P} \le \\ \varepsilon_t + \sum_{s=t+1}^{T-1} \|q(X)_s - q_{\mathcal{H}}(X)_s\|_{2,P} \le \varepsilon_t + \sum_{s=1}^{T-t-1} 2^{s-1}\varepsilon_{t+s}. \tag{4.2}$$

*Proof.*

$$E_t = \|q(X)_t - q_{\mathcal{H}}(X)_t\|_{2,P} = \|q(X)_t - \mathrm{pr}_{\mathcal{H}_t}\big(y_{q_{\mathcal{H}},t+1}(X)\big)\|_{2,P} \le$$

$$\|q(X)_t - \mathrm{pr}_{\mathcal{H}_t}\big(y_{q,t+1}(X)\big)\|_{2,P} + \|\,\mathrm{pr}_{\mathcal{H}_t}\big(y_{q,t+1}(X) - y_{q_{\mathcal{H}},t+1}(X)\big)\|_{2,P} = I_1 + I_2.$$

By assumption (4.1)

$$I_1 = \|q(X)_t - \mathrm{pr}_{\mathcal{H}_t} P\big(y_{q,t+1}(X) \mid \mathcal{F}_t\big)\|_{2,P} \le \varepsilon.$$

For the second term a recursive application of proposition 2.3 yields

$$I_2 \le \|P\big(y_{q,t+1}(X) - y_{q_{\mathcal{H}},t+1}(X) \mid \mathcal{F}_t\big)\|_{2,P} \le \|q(X)_{t+1} - q_{\mathcal{H}}(X)_{t+1}\|_{2,P} + \\ \|P\big(y_{q,t+2}(X) - y_{q_{\mathcal{H}},t+2}(X) \mid \mathcal{F}_{t+1}\big)\|_{2,P} \le \sum_{s=t+1}^{T-1} \|q(X)_s - q_{\mathcal{H}}(X)_s\|_{2,P}.$$

The last estimate in (4.2) is obtained by observing that $E_{T-1} = \varepsilon_{T-1}$, $E_{T-2} = \varepsilon_{T-2} + \varepsilon_{T-1}$ and the recursion $E_t = \varepsilon_t + \varepsilon_{t+1} + 2\sum_{s=t+2}^{T-1} E_s$. □

It is well-known that the rate of approximation depends on the degree of smoothness of the function to be approximated. Under additional assumptions on the Markov transition functions and the payoff function, theorem 4.1 can be strengthened to provide such a rate of convergence. Even though $v_t = \max(f_t, q_t)$ has only one weak derivative, the $q$-function

$$q_t(x) = P_{t,t+1}\big(d_{t,t+1}v_{t+1}\big)(x) = \int_{\mathbb{R}^d} d_{t,t+1}(y)v_{t+1}(y)p_{t,t+1}(x, dy) \tag{4.3}$$

is typically infinitely differentiable, because $P_{t,t+1} = P(. \mid \sigma(X_t))$ exhibits strong smoothing properties.

### 4.2. **Polynomial Approximation.**

Let $\mathcal{P}_m$ be the space of multivariate polynomials on $\mathbb{R}^d$ and coordinatewise degree at most $m-1$. Let $\mathcal{O} \subset \mathbb{R}^d$ be a smooth bounded domain and denote by $W_p^k(\mathcal{O})$ the usual Sobolev space with weak derivatives in $L_p(\mathcal{O})$ up to order $k$. The following is a classical result in approximation theory, see for example [29].

**Theorem 4.2.** *Let $1 \le p \le \infty$. Then, for any $f \in W_p^k(\mathcal{O})$*

$$\inf_{p\in\mathcal{P}_m} \|f-p\|_{L_p(\mathcal{O})} \le C\, m^{-k}\, \|f\|_{W_p^k(\mathcal{O})} \le C \dim(\mathcal{P}_m)^{-k/d}\, \|f\|_{W_p^k(\mathcal{O})}, \tag{4.4}$$

*where the constant $C$ is independent of $f$ and $m$.*

Polynomial approximation of functions on the whole Euclidian space $\mathbb{R}^d$ is more involved and necessitates weighted norms. Let $w(x)$ be a weight function on $\mathbb{R}^d$. Introduce the spaces $L_{p,w}(\mathbb{R}^d)$ of functions $f$ with weighted $p$-norm

$$\int |f(x)w(x)|^p dx < \infty$$

and let $W_{p,w}^k(\mathbb{R}^d)$ be the weighted Sobolev space with weak derivatives in $L_{p,w}(\mathbb{R}^d)$ up to order $k$.

The following theorem is a multivariate extension of the univariate weighted approximation results of [16] for special tensor product weights of Freud type. A survey of univariate weighted polynomial approximation can be found in [30], an extensive presentation of the whole theory is available in [31].

**Theorem 4.3.** *Let $1 \le p \le \infty$ and*

$$w_{\alpha,\beta}(x) = \prod_{i=1}^d \exp\big(-\beta|x_i|^\alpha\big), \tag{4.5}$$

*for $\alpha \ge 2$. Then, for any $f \in W_{p,w_{\alpha,\beta}}^k(\mathbb{R}^d)$*

$$\inf_{p\in\mathcal{P}_m} \|f-p\|_{L_{p,w_{\alpha,\beta}}(\mathbb{R}^d)} \le \\ C\, m^{-k(1-\frac{1}{\alpha})}\, \|f\|_{W_{p,w_{\alpha,\beta}}^k(\mathbb{R}^d)} \le C \dim(\mathcal{P}_m)^{-k(1-\frac{1}{\alpha})/d}\, \|f\|_{W_{p,w_{\alpha,\beta}}^k(\mathbb{R}^d)}, \tag{4.6}$$

*where $C$ is a universal constant, independent of $f$ and $m$, only dependent on the weight $w_{\alpha,\beta}$.*

Not much is published on multivariate weighted approximation. One reason is that tensor product approximation results are somewhat routine to obtain, given the univariate results [32]. For completeness we briefly sketch how to use the univariate theory to derive multivariate results.

*Proof.* The constant $c$ denotes a universal constant, which might be different wherever it appears. Let

$$E_{p,m}(f) = \inf_{p \in \mathcal{P}_m} \|f - p\|_{L_{p,w}(\mathbb{R})}.$$

A central tool is the shifted average operator $v_m$, defined in [31, equation (3.4.4)]. If $f \in L_{p,w}(\mathbb{R})$ then $v_m(f)$ is a polynomial of degree $2m-1$ and

$$E_{p,2m-1}(f) \leq \|f - v_m(f)\|_{L_{p,w}(\mathbb{R})} \leq cE_{p,m}(f).$$

The key estimates are the so called Jackson-Favard estimates, which state that for every differentiable function $f \in L_{p,w}(\mathbb{R})$

$$E_{p,m}(f) \leq c\frac{q_m}{m}E_{p,m-1}(f'). \tag{4.7}$$

For so called Freud weights [31, definition 3.1.1] the Freud numbers $q_m$ are determined as the least positive solution of $q_m Q'(q_m) = m$, where $Q(x) = \log(w(x)^{-1})$, and satisfy

$$c_1 q_m \leq q_{2m} \leq c_2 q_m \tag{4.8}$$

for constants $c_1, c_2$ only depending on $w$. For Freud weights of the concrete type $w(x) = w_{\alpha,\beta}(x) = \exp(-\beta|x|^\alpha)$

$$q_m = \left(\frac{m}{\alpha\beta}\right)^{\frac{1}{\alpha}}. \tag{4.9}$$

These results can be found in [31, section 3,4] or [30, theorem 4]. Assume $f$ is $k$-times differentiable with $\partial^k f \in L_{p,w}(\mathbb{R})$. Iterating (4.7) and applying the obvious bound $E_{p,m-k}(\partial^k f) \leq \|\partial^k f\|_{L_{p,w}(\mathbb{R})}$ shows

$$\|f - v_m(f)\|_{L_{p,w}(\mathbb{R})} \leq E_{p,m}(f) \leq c\left(\frac{q_m}{m}\right)^k E_{p,m-k}(\partial^k f) \leq \\ c(\alpha,\beta)m^{-k+1/\alpha}\|\partial^k f\|_{L_{p,w}(\mathbb{R})}, \tag{4.10}$$

which proves the result in one dimension. For higher dimensions $d$ introduce the multivariate shifted average operators $v_m^{[d]}(f)$ as in [31, equation (11.2.8)]: Denote by $v_{m,j}(f)$ the operator $v_m$ applied to $f$ as a function of the $j$-th coordinate. Then define $v_m^{[0]}(f) = f$, $v_m^{[1]}(f) = v_{m,1}(f)$ and $v_m^{[j]}(f) = v_m^{[j-1]}(v_{m,j}(f))$. A telescoping argument yields

$$\|f - v_m^{[d]}(f)\|_{L_{p,w}(\mathbb{R}^d)} \leq \sum_{j=1}^d \|v_m^{[j-1]}(f - v_{m,j}(f))\|_{L_{p,w}(\mathbb{R}^d)} \leq \\ c\sum_{j=1}^d \|f - v_{m,j}(f)\|_{L_{p,w}(\mathbb{R}^d)},$$

because all $v_m^{[j]}$ are bounded linear operators [31, theorem 3.4.2]. Using the product structure of the weights and Fubini, the summands on the right hand side can be estimated by the univariate estimate (4.10) such that for all $i \leq k$,

$$\|f - v_{m,j}(f)\|_{L_{p,w}(\mathbb{R}^d)} \leq c\left(\frac{q_m}{m}\right)^i E_{p,m-i}(\partial_j^i f).$$

To derive the final result (4.6) note that $v_m^{[d]}(f)$ is a polynomial of coordinatewise degree at most $2m-1$ such that

$$\inf_{p \in \mathcal{P}_{2m}} \|f - p\|_{L_{p,w_{\alpha,\beta}}(\mathbb{R}^d)} \leq \|f - v_m^{[d]}(f)\|_{L_{p,w}(\mathbb{R}^d)},$$

and apply the estimate $c_1 q_m \leq q_{2m}$ followed by the explicit representation (4.9). The generalization from smooth functions to functions in the Sobolev space $W^k_{p,w_{\alpha,\beta}}(\mathbb{R}^d)$ is standard. □

*Remark* 4.4. We would like to stress that the proof is constructive. The polynomial $v_m^{[d]}(f)$ is a linear combination in tensor products of Freud polynomials introduced in [31, equation (3.1.12a)], which are special orthogonal polynomials with respect to the weighted scalar product $L_{2,w}(\mathbb{R}^d)$.

Note that both bounds depend on the dimension of the state space. To evade the curse of dimensionality if the dimension of the state space increases, the degree of smoothness has to be increased in order to keep the rate of convergence constant.

Theorems 4.2 and 4.3 can be applied to derive the approximation rate for specific cases.

**Corollary 4.5.** *Consider the linear approximation architecture $\mathcal{H}_t = \mathcal{P}_m$. Let $\mathcal{O}$ be a smooth domain in $\mathbb{R}^d$. Assume that $f \in L_p(X)$ for some $p \geq 2$, that*

$$r_{\mathcal{O}}\, P_{t,t+1} : L_p(\mathbb{R}^d, \mu_{X_t}) \to W_2^k(\mathcal{O}), \tag{4.11}$$

*and that for some constant $c > 0$,*

$$\|1_{\mathcal{O}}\, g\|_{2,\mu_{X_t}} \leq c\|g\|_{L_2(\mathcal{O})} \quad \forall g \in L_2(\mathbb{R}^d, \mu_{X_t}). \tag{4.12}$$

*Then*

$$\|1_{\mathcal{O}}\,(q_t - q_{\mathcal{H},t})\|_{2,\mu_{X_t}} \leq c\|q_t - q_{\mathcal{H},t}\|_{L_2(\mathcal{O})} \leq C\, 2^{T-t-1} \dim(\mathcal{P}_m)^{-k/d} \sum_{s=t}^{T-1} \|q_s\|_{W_2^k(\mathcal{O})} \tag{4.13}$$

*for a constant $C$ independent of $q$, $m$.*

**Corollary 4.6.** *Consider the linear approximation architecture $\mathcal{H}_t = \mathcal{P}_m$. Assume that $f \in L_p(X)$ for some $p \geq 2$, that*

$$P_{t,t+1} : L_p(\mathbb{R}^d, \mu_{X_t}) \to W^k_{2,w_{2,\beta}}(\mathbb{R}^d) \tag{4.14}$$

*for some $k > 0$, $\beta > 0$, and that for some constant $c > 0$,*

$$\|g\|_{2,\mu_{X_t}} \leq c\|g\|_{L_{2,w_{2,\beta}}(\mathbb{R}^d)} \quad \forall g \in L_2(\mathbb{R}^d, \mu_{X_t}). \tag{4.15}$$

*Then*

$$\|q_t - q_{\mathcal{H},t}\|_{2,\mu_{X_t}} \leq c\|q_t - q_{\mathcal{H},t}\|_{L_{2,w_{2,\beta}}(\mathbb{R}^d)} \leq C\, 2^{T-t-1} \dim(\mathcal{P}_m)^{-k/(2d)} \sum_{s=t}^{T-1} \|q_s\|_{W^k_{2,w_{2,\beta}}(\mathbb{R}^d)} \tag{4.16}$$

*for a constant $C$ independent of $q$, $m$.*

*Proof.* Both corollaries follow from theorem 4.1 combined with theorems 4.2 respectively 4.3. □

In proposition 2.1, the error of truncating the payoff function has been estimated. In the following example we therefore deliberately restrict ourselves to bounded payoff functions.

*Example* 4.7. Let $P_{t,t+1}$ be a convolution operator

$$P_{t,t+1}(g) = g * K_t = \int_{\mathbb{R}^d} g(y)K_t(x-y)dy \tag{4.17}$$

for some kernel $K_t \in C_b^m(\mathbb{R}^d)$ with $\partial_\alpha K_t \in L_1(\mathbb{R}^d)$ for all $|\alpha| \leq k$. Then for every $g \in L_\infty(\mathbb{R}^d)$, $P_{t,t+1}(g) \in C_b^k(\mathbb{R}^d)$ and

$$\partial_\alpha P_{t,t+1}(g) = g * \partial_\alpha K_t \tag{4.18}$$

for all $|\alpha| \leq k$. Obviously, $M_{1_\mathcal{O}} : C_b^k(\mathbb{R}^d)) \to W_2^k(\mathcal{O})$ and $C_b^k(\mathbb{R}^d) \subset W_{p,w_{2,\beta}}^k(\mathbb{R}^d)$ for any $\beta > 0$. Hence conditions (4.11) and (4.14) are satisfied for $p = \infty$.

If $X_t$ is sampled a time steps of width $\Delta t$ from a diffusion at with uniformly elliptic generator, the transition functions satisfy the Gaussian bounds

$$K_t(x) \leq c(\Delta t)^{-d/2} \exp\big(-\frac{\lambda}{\Delta t}\|x\|^2\big) \tag{4.19}$$

for some constants $c, \lambda$. Consequently, conditions (4.12) and (4.15) are satisfied as well.

Note that example 4.7 covers the traditional multivariate Black-Scholes model, where $X_t = (\log S_{1,t}, \ldots, \log S_{d,t})$.

## 5. Review of Relevant Results from Empirical Process Theory

Before we proceed to the analysis of the sample error we give an excerpt of the key results on empirical process theory which are needed later on. A thorough development of the topic can be found in the books of Pollard [35, chapter II], [36], Dudley [13], or van der Vaart and Wellner [42].

Consider a sequence of iid random variables $X_1, \ldots, X_n, \ldots$ on a probability space $(\Omega, P, \mathcal{F})$. Introduce the empirical measure

$$P_n f = \frac{1}{n}\sum_{i=1}^n \delta_{X_i} f = \frac{1}{n}\sum_{i=1}^n f(X_i) \tag{5.1}$$

where $\delta_x$ is the Dirac measure concentrated at $x$, and the empirical process

$$E_n = \sqrt{n}(P_n - P). \tag{5.2}$$

The ordinary strong Law of Large Numbers states that for a function $f \in L_1(\Omega, P, \mathcal{F})$

$$\lim_{n\to\infty} |P_n f - Pf| = 0$$

almost sure, and if $P(f^2) < \infty$,

$$\sqrt{n}(P_n f - Pf) \to N(0, \operatorname{Var} f)$$

in distribution, by the ordinary Central Limit Theorem. For many applications, in particular for empirical risk minimization as we have seen in the introduction, these point-wise convergence results are not sufficient. The objective of empirical process theory is to extend them uniformly to whole classes of functions. Uniform Law of Large Numbers provide conditions on a class of functions $\mathcal{G} \subset \mathcal{L}_1(\Omega, P, \mathcal{F})$ and on the underlying probability measure such that

$$\lim_{n\to\infty} \sup_{f\in\mathcal{G}} |P_n f - Pf| = 0$$

almost sure. The earliest result in this direction is the well-known Glivenko-Cantelli theorem, for which the class $\mathcal{G}$ are the indicator functions $1_{(\infty,t]}$, $t \in \mathbb{R}$. In their seminal paper [44], Vapnik and Chervonenkis proved that classes of sets satisfying a combinatorial condition, nowadays called VC-classes, satisfy a uniform Law of Large Numbers, to be precise they showed convergence in probability. Similarly, the uniform Central Limit Theorems provide conditions, such that the empirical process $E_n = \sqrt{n}(P_n - P)$, indexed by $f \in \mathcal{G} \subset \mathcal{L}_2(\Omega, P, \mathcal{F})$, converges weakly in $l_\infty(\mathcal{G})$ to a Brownian bridge.

To properly deal with empirical measure and the empirical process as random elements, introduce a countable product space

$$(\Omega^\infty, \mathbb{P}, \mathcal{F}^\infty), \tag{5.3}$$

where $\mathbb{P} = P^{\otimes\infty}$ is the product measure and $\mathcal{F}^\infty$ is the product $\sigma$-algebra. The random variable $X_i$ can now be identified with the $i$-th coordinate projection.

### 5.1. Entropy and Covering Numbers.

Let $(\mathcal{M}, d)$ be a semi-metric space[2]. A subset $\{x_1, \ldots x_n\} \subset \mathcal{M}$ is called $\varepsilon$-net if $\forall\, x \in \mathcal{M}$, $\exists\, x_i$, such that $d(x, x_i) \le \varepsilon$. Define the covering number by

$$N(\varepsilon, \mathcal{M}, d) = \inf\{n \in \mathbb{N} \mid \exists\ \varepsilon\text{-net } \{x_1, \ldots x_n\} \text{ of cardinality } n\}. \tag{5.4}$$

It is the minimum number of closed balls of radius $\varepsilon$ required to cover $\mathcal{M}$. A quantity related to the covering number is the packing number $D(\varepsilon, \mathcal{M}, d)$, defined as the maximal number of disjoint $\varepsilon$-balls that can be packed into $\mathcal{M}$. Then

$$D(2\varepsilon, \mathcal{M}, d) \le N(\varepsilon, \mathcal{M}, d) \le D(\varepsilon, \mathcal{M}, d). \tag{5.5}$$

If follows directly from the definition that if two metrics satisfy $d_2 \le c d_1$ for some constant $c > 0$, then

$$N(\varepsilon, \mathcal{M}, d_2) \le N(\varepsilon, \mathcal{M}, c\, d_1) = N(\frac{\varepsilon}{c}, \mathcal{M}, d_1). \tag{5.6}$$

The function $H(\cdot, \mathcal{M}, d) = \log N(\cdot, \mathcal{M}, d)$ is called the metric entropy of $(\mathcal{M}, d)$.

*Example* 5.1. For all $\delta < R$ the $\delta$-covering number of a ball $B_R$ of radius $R$ in the Euclidian space $\mathbb{R}^n$ is bounded by

$$N(\delta, B_R, d_e) \le \left(\frac{4R}{\delta}\right)^n. \tag{5.7}$$

Let $(\Omega, \mathcal{F})$ be a measure space and $P$ a probability measure on $(\Omega, \mathcal{F})$. Let $\mathcal{L}_p(\Omega, \mathcal{F}, P)$ denote the set of all $p$-integrable functions with respect to the measure $P$ and $L_p(\Omega, \mathcal{F}, P)$ the corresponding space of equivalence classes. If the measure space is clear from the context or not relevant we just drop it in the notation. The spaces $\mathcal{L}_p(P)$ are endowed with the usual semi-metric $d_{p,P}(f, g) = \|f - g\|_{p,P}$. A measurable function $G$ is called an envelope of $\mathcal{G}$ if $|g| \le G$ for every $g \in \mathcal{G}$. Finally define

$$N_p(\varepsilon, \mathcal{G}) = \sup_P N(\varepsilon, \mathcal{G}, d_{p,P}), \tag{5.8}$$

where the supremum runs over all over all probability measures $P$ on $(\Omega, \mathcal{F})$ concentrated in finite sets.

The following lemmas are useful to bound the covering numbers.

**Lemma 5.2.** *Let $(\mathcal{M}_i, d_i)$ be metric spaces and $F : \mathcal{M}_1 \to \mathcal{M}_2$ a surjective Lipschitz map with $d_2(F(x_1), F(x_2)) \le C d_1(x_1, x_2)$. Then*

$$N(C\varepsilon, \mathcal{M}_2, d_2) \le N(\varepsilon, \mathcal{M}_1, d_1).$$

*Proof.* This follows directly from (5.6). □

[2] Possibly $d(x_1, x_2) = 0$ for some $x_1 \neq x_2$.

**Lemma 5.3.** *Let $\mathcal{F}$ and $\mathcal{G}$ be classes of measurable functions with envelopes $F$ respectively $G$. Then for every operation $\star \in \{+, -, \wedge, \vee\}$ and constants $a, b > 0$,*

$$\begin{aligned}
N(\varepsilon(a+b), \mathcal{F} \star \mathcal{G}, d_{p,P}) &\leq N(a\varepsilon, \mathcal{F}, d_{p,P})\, N(b\varepsilon, \mathcal{G}, d_{p,P}), && (5.9)\\
N(\varepsilon(a+b), \mathcal{F} \cdot \mathcal{G}, d_{p,P}) &\leq N(a\varepsilon, \mathcal{F}, d_{p,G^p \cdot P})\, N(b\varepsilon, \mathcal{G}, d_{p,F^p \cdot P}), && (5.10)\\
N(\varepsilon, \mathcal{F}^2, d_{p,P}) &\leq N(\varepsilon/2, \mathcal{F}, d_{p,F^p \cdot P}), && (5.11)\\
N(\varepsilon, \mathcal{F}, d_{p,P}) &\leq N(\varepsilon^p, \mathcal{F}, d_{1,(2F)^{p-1} \cdot P}), && (5.12)\\
N(\varepsilon, \mathcal{F}, d_{p,F \cdot P}) &\leq N(\varepsilon/\|F\|_{r,P}^{1/p}, \mathcal{F}, d_{p\,q,P}), && (5.13)
\end{aligned}$$

*where $\mathcal{F} * \mathcal{G} = \{f * g \mid f \in \mathcal{F}, g \in \mathcal{G}\}$ for $* \in \{\cdot, +, -, \wedge, \vee\}$, $\mathcal{F}^2 = \mathcal{F} \cdot \mathcal{F}$, and $F \cdot P$ is the measure $F \cdot P(A) = P(1_A F)$ and $1/q + 1/r = 1$ are conjugate exponents.*

*Proof.* Let $\{f_i\} \subset \mathcal{F}$ and $\{g_i\} \subset \mathcal{G}$ appropriate minimal nets. The first relation follows from the triangle inequality. For (5.10)

$$\begin{aligned}
P(|fg - g_i f_i|^p)^{1/p} \leq P(|fg_i - f_i g_i|^p)^{1/p} + P(|fg - fg_i|^p)^{1/p} \leq \\
P(|f - f_i|^p G^p)^{1/p} + P(|g - g_i|^p F^p)^{1/p}.
\end{aligned}$$

As for (5.11)

$$P(|f^2 - f_i^2|^p)^{1/p} = P(|f + f_i|^p |f - f_i|^p)^{1/p} \leq 2P(F^p |f - f_i|^p)^{1/p}.$$

Inequality (5.12) follows similarly

$$P(|f - f_i|^p)^{1/p} = P(|f - f_i||f - f_i|^{p-1})^{1/p} \leq P(|f - f_i|(2F)^{p-1})^{1/p}.$$

Finally, (5.13) is a consequence of Hölder's inequality. □

A typical application of the above lemma is to convert the $\mathcal{L}_2$ covering numbers into $\mathcal{L}_1$ covering numbers. (5.11) and (5.13) imply

$$N(\varepsilon, \mathcal{F}^2, d_{1,P}) \leq N(\varepsilon/2, \mathcal{F}, d_{1,F \cdot P}) \leq N(\varepsilon/(2\|F\|_{2,P}), \mathcal{F}, d_{2,P}), \tag{5.14}$$

or

$$N(2\varepsilon\|F^2\|_{1,P}, \mathcal{F}^2, d_{1,P}) \leq N(\varepsilon\|F\|_{2,P}, \mathcal{F}, d_{2,P}). \tag{5.15}$$

5.2. **Vapnik-Chervonenkis Classes.** Let $\mathcal{C}$ be a class of subsets of an arbitrary set $\mathcal{X}$. A set $x = (x_1, \dots, x_n) \in \mathcal{X}^n$ of $n$ points is shattered by $\mathcal{C}$ if

$$\Delta_{\mathcal{C}}(x) = |\{C \cap \{x_1, \dots, x_n\} \mid C \in \mathcal{C}\}| = 2^n. \tag{5.16}$$

In terms of indicator functions, $x = (x_1, \dots, x_n)$ is shattered by $\mathcal{C}$ if

$$\{\big(1_C(x_1), \dots, 1_C(x_n)\big) \mid g \in \mathcal{G}\} = \{0,1\}^n. \tag{5.17}$$

The VC dimension $\dim_{VC}(\mathcal{C})$ of $\mathcal{C}$ is the cardinality of the largest discrete subset of $\mathcal{X}$ shattered by $\mathcal{C}$

$$\dim_{VC}(\mathcal{C}) = \sup\{n \mid \exists\, x \in \mathcal{X}^n \text{ s.t. } \Delta_{\mathcal{C}}(x) = 2^n\}. \tag{5.18}$$

The class $\mathcal{C}$ is a VC class if $\dim_{VC}(\mathcal{C}) < \infty$. A peculiar property of a VC class is that the growth function

$$\Delta_{\mathcal{C}}(n) = \max_{x \in \mathcal{X}^n} \Delta_{\mathcal{C}}(x) \tag{5.19}$$

is polynomial in $n$, more precisely, if $\dim_{VC}(\mathcal{C}) = d$,

$$\Delta_{\mathcal{C}}(n) \leq \phi(n, d), \tag{5.20}$$

for $n \geq d \geq 1$, where, by Stirling's formula,

$$\phi(n, d) = \sum_{i=0}^{d} \binom{n}{i} \leq 1.5 \frac{n^d}{d!} \leq \Big(\frac{e\, n}{d}\Big)^d, \tag{5.21}$$

denotes the number of subsets of a set of cardinality $n$, which contain at most $d$ elements. (5.20) is known as Sauer's lemma, the first bound in (5.21) is due to Vapnik and Chervonenkis. VC classes have a variety of permanence properties which allow the construction of new VC classes form basic VC classes by simple operations such as complements, intersections, unions or products.

The covering number of VC classes can only grow at a polynomial rate. Let $(\Omega, P, \mathcal{F})$ be a probability space and introduce on $\mathcal{F}$ the pseudo-metric

$$d_P(A, B) = P(A\Delta B),$$

where $\Delta$ is the symmetric difference. Haussler improved the $\mathcal{L}_1$ packing number bound of Dudley [12] by a logarithmic factor.

**Theorem 5.4** (Haussler [18])**.** *Let $\mathcal{X}$ be a set. For any probability measure $P$ on $\mathcal{X}$, any class $\mathcal{C}$ of $P$-measurable sets of* $\dim_{VC} = d < \infty$ *and any $\varepsilon > 0$,*

$$N(\varepsilon, \mathcal{C}, d_P) \leq D(\varepsilon, \mathcal{C}, d_P) \leq e(d+1)\left(\frac{2e}{\varepsilon}\right)^d. \tag{5.22}$$

It is worthwhile to remark that Haussler established also lower bounds on the packing number in [18].

Let $\mathcal{G}$ be a class of real-valued functions on $\mathcal{X}$. A set $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ is pseudo-shattered by $\mathcal{G}$, if there are real numbers $t = (t_1, \ldots, t_n)$ such that

$$\{\big(\operatorname{sign}(g(x_1) - t_1), \ldots, \operatorname{sign}(g(x_n) - t_n)\big) \mid g \in \mathcal{G}\} = \{0, 1\}^n, \tag{5.23}$$

in other words, given a fixed translation vector $t$, every boolean vector $b \in \{0, 1\}^n$ can be realized by a function $g \in \mathcal{G}$. The pseudo-dimension of $\mathcal{G}$ is the largest integer $n$ such that there exists a set of cardinality $n$ which is pseudo-shattered by $\mathcal{X}$, or

$$\dim_P(\mathcal{G}) = \sup\{n \mid \exists\, x \in \mathcal{X}^n \text{ pseudo-shattered by } \mathcal{G}\}. \tag{5.24}$$

The pseudo-dimension has been formulated by [36] and [18]. It can also be introduced via the concept of subgraph classes. The subgraph of a real-valued function $f$ on an arbitrary set $\mathcal{X}$ is defined as

$$G_f = \{(x, t) \in \mathcal{X} \times \mathbb{R} \mid t \leq f(x)\}. \tag{5.25}$$

A class of real-valued functions $\mathcal{G}$ on $\mathcal{X}$ is called a VC subgraph class if its class of subgraphs is a VC class. Because a set is shattered by the subgraph class $\{G_f \mid f \in \mathcal{G}\}$ if and only if it is pseudo-shattered by the class of indicator functions $\{\operatorname{sign}(f(x) - t) \mid f \in \mathcal{G}\}$, the pseudo-dimension of $\mathcal{G}$ is equal to the VC dimension of its subgraph class.

**Proposition 5.5.** *Let $\mathcal{G}$ be a finite dimensional real vector space of measurable real-valued functions. Then the class of sets $nn(\mathcal{G}) = \{\{g \geq 0\} \mid g \in \mathcal{G}\}$ is a VC class with* $\dim_{VC}(nn(\mathcal{G})) = \dim(\mathcal{G})$*. If $f$ is a fixed function, then* $\dim_{VC}(nn(f + \mathcal{G})) = \dim_{VC}(nn(\mathcal{G}))$*. Finally, $\mathcal{G}$ is a VC subgraph class with finite pseudo-dimension* $\dim_P(\mathcal{G}) = \dim(\mathcal{G})$*.*

*Proof.* For the first two statements we refer to [13, theorem 4.2.1]. The last statements follows from the first two: Consider the affine class of functions $f + \mathcal{G}$ on $\mathcal{X} \times \mathbb{R}$, where $f(x, t) = -t$ and note that the subgraph class of $\mathcal{G}$ is precisely $nn(f + \mathcal{G})$. □

Theorem 5.4 can be used to bound the $\mathcal{L}_p$ covering number of classes of functions with finite pseudo-dimension. A proof of the following result can also be found in [42, theorem 2.6.7], however, the constants are not quite correct.

**Theorem 5.6.** *Let $\mathcal{G}$ be a class of measurable functions with finite pseudo-dimension* $\dim_P(\mathcal{G}) = d < \infty$ *and measurable envelope $G$. Then, for $p \geq 1$ and any probability measure $P$ with $\|G\|_{p,P} > 0$,*

$$N(\varepsilon\|G\|_{p,P}, \mathcal{G}, d_{p,P}) \leq e(d+1)\left(\frac{2^{p+1}e}{\varepsilon^p}\right)^d \tag{5.26}$$

*for all* $0 < \varepsilon < 1$.

*Proof.* As in [18] or [42] note that by Fubini, $P|f-g| = (P \times \lambda)(G_f \Delta G_g)$ where $\lambda$ is the Lebesgue measure on $\mathbb{R}$. Normalize $P \times \lambda$ to a probability measure $Q = (P \times \lambda)/(2PG)$ on $\{(x,t) \in \mathcal{X} \times \mathbb{R} \mid |t| \leq G(x)\}$. Theorem 5.4 implies for any probability measure $P$

$$N(\varepsilon\, 2\, PG, \mathcal{G}, d_{1,P}) = N(\varepsilon\, 2\, PG, G_{\mathcal{G}}, d_{P \times \lambda}) = N(\varepsilon, G_{\mathcal{G}}, d_Q) \leq e(d+1)\left(\frac{2e}{\varepsilon}\right)^d,$$

where $G_{\mathcal{G}}$ is the class of subgraphs of $\mathcal{G}$. As for $p > 1$ apply (5.12)

$$N(\varepsilon, \mathcal{G}, d_{p,P}) \leq N(\varepsilon^p, \mathcal{G}, d_{1,(2G)^{p-1}\cdot P}) = N(\varepsilon^p \frac{QG}{2^{p-1}P(G^p)}, \mathcal{G}, d_{1,Q}),$$

where $Q = G^{p-1}/P(G^{p-1}) \cdot P$ is the normalization of $G^{p-1} \cdot P$ to a probability measure. Apply these two estimates to conclude

$$N(\varepsilon \|G\|_{p,P}, \mathcal{G}, d_{p,P}) \leq N((\varepsilon/2)^p\, 2\, QG, \mathcal{G}, d_{1,Q}) \leq e(d+1)\left(\frac{2^{p+1}e}{\varepsilon^p}\right)^d.$$

□

### 5.3. **Uniform Law of Large Numbers.** The generalized Glivenko-Cantelli theorem can now be stated.

**Theorem 5.7** (Uniform Law of Large Numbers, [13, theorem 6.1.7])**.** *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space and* $\mathcal{G}$ *a class of measurable functions on* $\Omega$ *with envelope* $G \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$. *Assume that* $\mathcal{G}$ *satisfies appropriate measurability conditions. If*

$$N_1(\varepsilon, \mathcal{G}) = \sup_Q N(\varepsilon, \mathcal{G}, d_{1,Q}) < \infty \quad \forall\, \varepsilon > 0, \tag{5.27}$$

*where the supremum runs over all probability measures* $Q$ *concentrated on finite sets, then*

$$\lim_{n\to\infty} \sup_{g \in \mathcal{G}} |P_n g - Pg| = 0 \qquad \textit{almost sure.} \tag{5.28}$$

*Remark* 5.8. The condition (5.27) can be replaced in fact by

$$\log N(\varepsilon, \mathcal{G}, d_{1,P_n}) = o_P(n), \tag{5.29}$$

which is, if $PG < \infty$, equivalent to

$$\log N(\varepsilon \|G\|_{1,P_n}, \mathcal{G}, d_{1,P_n}) = o_P(n). \tag{5.30}$$

To avoid measurability problems which might be caused by taking the supremum over an uncountable class $\mathcal{G}$, (5.27) is normally formulated in terms of an outer measure. However, a rather weak condition called *image admissible Suslin* is sufficient to rule out any measurability problems. See for example [13, chapter 5].

**Definition 5.9.** A class of functions $\mathcal{G}$ for which the generalized Glivenko-Cantelli theorem 5.7 holds is called a Glivenko-Cantelli class (for $P$).

The covering number bounds (5.26) in theorem 5.6 show that for VC subgraph classes with $\mathcal{L}_1$-envelope, the left hand side of (5.30) is uniformly bounded for every fixed $\varepsilon$ and are therefore Glivenko-Cantelli.

5.4. **Uniform Deviation from the Mean.** The proof of theorem 5.7 relies on the following estimate due to Pollard [35, page 26], see also [11, theorem 29.1]. For a class $\mathcal{G}$ of measurable functions uniformly bounded by $K$

$$\mathbb{P}\big(\sup_{g\in\mathcal{G}} |P_n g - Pg| > \varepsilon\big) \le 8\mathbb{E}[N(\frac{\varepsilon}{8}, \mathcal{G}, d_{1,P_n})] \ \exp\left(-\frac{n\varepsilon^2}{128K^2}\right). \tag{5.31}$$

Remember that $\mathbb{P} = P^{\otimes\infty}$ denotes the product measure. Similar uniform deviation inequalities have been obtained earlier by Vapnik and Chervonenkis for VC classes of sets and VC-major classes. This estimate has been substantially improved by Talagrand in [39]. He considered functions with values in $[0,1]$. Let $\mathcal{G}$ be a class of measurable functions with range $[0,K]$ of pseudo-dimension $d$. Then by (5.26)

$$N(\varepsilon, \mathcal{G}/K, d_{2,P_n}) = N(\varepsilon K, \mathcal{G}, d_{2,P_n}) \le e(d+1)\left(\frac{\sqrt{8e}}{\varepsilon}\right)^{2d}. \tag{5.32}$$

A straightforward scaling argument combined with [39, theorem 1.3] leads to

$$\mathbb{P}\big(\sup_{g\in\mathcal{G}} |P_n g - Pg| > \varepsilon\big) \le \left(C(V)\frac{\sqrt{n}\,\varepsilon}{\sqrt{2d}\,K}\right)^{2d} \exp\left(-\frac{2n\varepsilon^2}{K^2}\right), \tag{5.33}$$

where $C(V)$ is a constant only depending on $V = (e(d+1))^{1/(2d)}\sqrt{8e}$. The bound (5.33) does not just holds for VC subgraph classes but whenever an entropy estimate of the form (5.32) holds. It also shows that the convergence rate of the expectation is

$$\mathbb{P}\sup_{g\in\mathcal{G}} |P_n g - Pg| \le O\left(\frac{1}{\sqrt{n}}\right), \tag{5.34}$$

whereas (5.31) would result in a rate of $O\big(\sqrt{\log n/n}\big)$. For other new deviation inequalities see also [14].

5.5. **Uniform Central Limit Theorems.** Consider a class $\mathcal{G}$ of functions endowed with a semi-metric $d_{\mathcal{G}}$. The class $\mathcal{G}$ is said to be asymptotically equicontinuous if for every $\varepsilon > 0$,

$$\lim_{\delta\to 0}\limsup_{m\to\infty} \mathbb{P}\big(\sup_{d_{\mathcal{G}}(f,g)<\delta} \mid E_n(f-g) \mid > \varepsilon\big) = 0. \tag{5.35}$$

Same remarks on measurability as for the generalized Glivenko-Cantelli theorem apply. For $\mathcal{G} \subset \mathcal{L}_2(P)$ the natural semi-metric to consider is

$$d_{\mathcal{G}}(f,g) = \rho_P(f-g), \tag{5.36}$$

where $\rho_P(f) = (P(f-Pf)^2)^{1/2}$. Let $l_\infty(\mathcal{G})$ be the metric space of all bounded functions $H$ from $\mathcal{G}$ to $\mathbb{R}$ endowed with the supremum norm $\sup_{\mathcal{G}} |H(f)|$. Note the difficulty that $l_\infty(\mathcal{G})$ is typically not separable.

**Definition 5.10.** A class of functions $\mathcal{G}$ is called a Donsker class for $P$ if it is pre-Gaussian and $E_n \Rightarrow B_P$ in $l_\infty(\mathcal{G})$. It is called uniformly Donsker if it is a Donsker class for every $P$.

Note that Donsker classes are Glivenko-Cantelli classes, but not conversely. In the above definition $B_P$ is the $P$-Brownian bridge. It is a Gaussian process indexed by $f \in \mathcal{G} \subset \mathcal{L}_2(P)$ with mean 0 and covariance function

$$P\big(B_P(f)B_P(g)\big) = P\big((f-Pf)(g-Pg)\big) = P(fg) - P(f)P(g).$$

The class $\mathcal{G}$ is pre-Gaussian if $B_P$ can be defined on a probability space such that for almost all $\omega$, $f \mapsto B_P(f)(\omega)$ is bounded and uniformly continuous for $\rho_P$ as a function from $\mathcal{G}$ to $\mathbb{R}$. Consequently $E_n$ and $B_P$ can be considered as random elements with values in the metric space $l_\infty(\mathcal{G})$. We refer to [13] for more details.

**Theorem 5.11** ([13, theorem 3.7.2])**.** *A class of measurable functions $\mathcal{G} \subset \mathcal{L}_2(P)$ is Donsker for $P$ if and only if $\mathcal{G}$ is totally bounded in $\mathcal{L}_2(P)$ and satisfies the asymptotic equicontinuity condition* (5.35) *with $d_{\mathcal{G}} = \rho_P$.*

We now require conditions to verify the Donsker property. This is achieved by imposing a uniform entropy bound.

**Theorem 5.12** (Koltchinskii-Pollard Central Limit Theorem, [13, theorem 6.3.1])**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{G} \subset \mathcal{L}_2(\Omega, \mathcal{F}, P)$ a class of functions with envelope $G \in \mathcal{L}_2(\Omega, \mathcal{F}, P)$. Assume that $\mathcal{G}$ satisfies appropriate measurability conditions. If*

$$\int_0^1 \sup_Q \sqrt{\log N(\varepsilon \|G\|_{2,Q}, \mathcal{G}, d_{2,Q})} d\varepsilon < \infty \tag{5.37}$$

*where the supremum is taken over probability measures $Q$ concentrated on finite sets, then $\mathcal{G}$ is Donsker for $P$.*

In light of the covering number bounds of VC subgraph classes we have

**Corollary 5.13.** *If $\mathcal{G} \subset \mathcal{L}_2(\Omega, \mathcal{F}, P)$ is a VC subgraph class with envelope $G \in \mathcal{L}_2(\Omega, \mathcal{F}, P)$, then $\mathcal{G}$ is Donsker for $P$.*

## 6. Sample Error

As in section 5, let $P_n$ be the empirical measure and $E_n$ the empirical process based on $n$ independent sample path $X^1, \dots, X^n$ of the Markov process $X$. Then, the approximate $q$-function $\hat{q}_{n,\mathcal{H}}$ can be viewed as a random element on $\Omega^\infty$ with values in $L_2(X)$ and the sample error as a random variable on $\Omega^\infty$. Assume from now on that the approximation architecture $\mathcal{H}$ is held fixed. Introduce the quadratic loss functions

$$l_t : L_2(\mathbb{R}^d, \mu_{X_t}) \times L_2(X) \to L_1(X), \quad l_t(u,h)(x) = \big(u(x_t) - y_{h,t+1}(x)\big)^2. \tag{6.1}$$

For $u, v, w \in L_2(\mathbb{R}^d, \mu_{X_t})$ and $h \in L_2(X)$, it follows that

$$l_t(u,h) = l_t(v,h) + (2v - y_{h,t+1})(u - v) + (u - v)^2, \tag{6.2}$$

showing that $l_t(u,h)$, as a function of $u$, is Fréchet differentiable with

$$D_1 l_t(u,h) = M_{(2u - y_{h,t+1})}, \quad D_1^2 l_t(u,h) = M_2. \tag{6.3}$$

Here, $M_a$ is the multiplication operator by $a$. Define the risk and empirical risk functional

$$\begin{aligned} L_t(u,h) &= P\, l_t(u,h)(X), \\ L_{t,n}(u,h) &= P_n\, l_t(u,h)(X). \end{aligned} \tag{6.4}$$

Integrating (6.2) shows that $L_t(u,h)$ is differentiable in $u$ with

$$\begin{aligned} D_1 L_t(u,h)(v) &= 2\,P\,\big(u(X_t) - y_{h,t+1}(X)\big) v(X_t), \\ D_1^2 L_t(u,h)(v,w) &= 2\,P\, v(X_t) w(X_t). \end{aligned} \tag{6.5}$$

Introduce the minimizers

$$\begin{aligned} q^*_{\mathcal{H},t}(h) &= \underset{u \in \mathcal{H}_t}{\arg\min}\, L_t(u,h), \\ q^*_{n,\mathcal{H},t}(h) &= \underset{u \in \mathcal{H}_t}{\arg\min}\, L_{n,t}(u,h). \end{aligned} \tag{6.6}$$

parameterized by $h \in L_2(X)$. Note that they only depend on $h_s$ for $s > t$. Consequently,

$$q_{\mathcal{H},t} = q^*_{\mathcal{H},t}(q_{\mathcal{H}}), \quad \hat{q}_{n,\mathcal{H},t} = q^*_{n,\mathcal{H},t}(\hat{q}_{n,\mathcal{H}}). \tag{6.7}$$

Fix $t$ and $h \in L_2(X)$ for the moment and let $\mathcal{H}$ be *any* approximation architecture. The inequalities (1.1) translate to

$$\begin{aligned} 0 \leq L_t(\hat{q}_n, h) - L_t(q^*, h) = L_t(\hat{q}_n, h) - \inf_{u\mathcal{H}_t} L_t(u,h) \leq \\ L_t(\hat{q}_n, h) - L_{t,n}(\hat{q}_n, h) + \sup_{u\in\mathcal{H}_t} \big(L_{t,n}(u,h) - L_t(u,h)\big) \leq \\ 2 \sup_{u\in\mathcal{H}_t} |L_{t,n}(u,h) - L_t(u,h)|, \end{aligned} \tag{6.8}$$

where $\hat{q}_n = q^*_{n,\mathcal{H},t}(h)$, $q^* = q^*_{\mathcal{H},t}(h)$, and

$$0 \leq L_{t,n}(\hat{q}_n, h) - L_t(\hat{q}_n, h) \leq \sup_{u\in\mathcal{H}_t} |L_{t,n}(u,h) - L_t(u,h)|. \tag{6.9}$$

Also note that from (6.2) and $D_1 L_t(q^*_{\mathcal{H},t}(h), h) = 0$

$$P\big(q^*_{n,\mathcal{H},t}(h)(X_t) - q^*_{\mathcal{H},t}(h)(X_t)\big)^2 = L_t(q^*_{n,\mathcal{H},t}(h), h) - L_t(q^*_{\mathcal{H},t}(h), h), \tag{6.10}$$

which provides (1.2). Almost sure convergence of the parameterized minimizers is now a direct consequence of a Uniform Law of Large Numbers as explained in the introduction.

6.1. **Almost Sure Convergence.** The application of empirical risk minimization to control the sample error of the Longstaff-Schwartz algorithm is slightly more involved. Instead of a fixed $h \in L_2(X)$ one has to deal with the approximate $q$-functions $\hat{q}_{n,\mathcal{H},s}$ of the previous time step $s = t{+}1, \ldots, T{-}1$, which is are themselves random elements.

Fix an approximation architecture $\mathcal{H} \subset L_2(X)$. Let $r > 0$ and introduce the following classes of functions

$$\begin{aligned} \mathcal{Y}_t &= \{y_{h,t} \mid h \in \mathcal{H}\}, \\ \mathcal{L}_t &= \{l_t(u,h) \mid u \in \mathcal{H}_t, h \in \mathcal{H}\}, \\ \mathcal{H}_t(r) &= \{u \in \mathcal{H}_t \mid \|u(X_t)\|_{2,P} \leq r\}, \\ \mathcal{L}_t(r) &= \{l_t(u,h) \mid u \in \mathcal{H}_t(r), h \in \mathcal{H}\}. \end{aligned} \tag{6.11}$$

The class $\mathcal{Y}_t$ has envelope $Y_t = \sum_{s=t+1}^T f_s \in L_2(\Omega, P, \mathcal{F})$. If $\mathcal{H}_t(r)$ is compact, which is for example the case whenever $\mathcal{H}$ is a linear approximation architecture of finite dimension, then $\mathcal{H}_t(r)$ and $\mathcal{L}_t(r)$ have an envelope as well.

The outline for proving convergence of the sample estimator $\hat{q}_{n,\mathcal{H}}$ is now as follows: Bound the covering number of the function classes (6.11) and prove a compactness result for the sample minimizers to derive almost sure convergence of $q^*_{n,\mathcal{H},t}(h)$. The continuity of the minimizers $q^*_{\mathcal{H},t}(h)$ in $h$ is then used to finalize the convergence proof.

**Proposition 6.1.** *Let $\mathcal{H}$ is a linear approximation architecture of $\dim(\mathcal{H}) = k$. Then, the class $\mathcal{H}_t$ has finite pseudo-dimension $dim_P(\mathcal{H}_t) = k$. Let $H_t$ be an envelop for $\mathcal{H}_t(r)$. Then, for any probability measure $Q$ on $(\Omega, \mathcal{F})$ with $\|H_t\|_{2,Q} > 0$, the covering number is bounded by*

$$N(\varepsilon\|H_t\|_{2,Q}, \mathcal{H}_t(r), d_{2,Q}) \leq e(k+1)\left(\frac{8e}{\varepsilon^2}\right)^k$$

*Proof.* Apply propositions 5.5 and 5.6. □

**Proposition 6.2.** *Let $\mathcal{H}$ is an approximation architecture of* $\dim_P(\mathcal{H}) = k$*. Then, the class $\mathcal{Y}_t$ has finite pseudo-dimension*

$$\begin{aligned} d_{\mathcal{Y}_t} = \dim_P(\mathcal{Y}_t) \leq \\ K \inf\{j \in \mathbb{N} \mid \frac{j}{\log_2(ej)} > (T-t)\} \leq 2(T-t)\log_2(e(T-t))k \end{aligned} \tag{6.12}$$

*In particular, for any probability measure $Q$ on $(\Omega, \mathcal{F})$ with $\|Y_t\|_{2,Q} > 0$, the covering number is bounded by*

$$N(\varepsilon\|Y\|_{2,Q}, \mathcal{Y}_t, d_{2,Q}) \leq e(d_{\mathcal{Y}_t} + 1)\left(\frac{8e}{\varepsilon^2}\right)^{d_{\mathcal{Y}_t}}.$$

*Proof.* First observe that if $\mathcal{G}$ is a vector space of real-valued functions of dimension $d$ and $g$ is an arbitrary real-valued function, then, by proposition 5.5, the class

$$nn(g + \mathcal{G}) = \{\{g + f \geq 0\} \mid f \in \mathcal{G}\}$$

is a VC class of VC dimension $d$, equivalently, the class of corresponding indicator functions has pseudo-dimension $d$. This applies in particular to the classes

$$\mathcal{A}_{\mathcal{H},t} = \big\{A_{h,t} = \{x \in \mathbb{R}^d \mid h(x)_t \leq f(x)_t\} \mid h \in \mathcal{H}\big\} \tag{6.13}$$

whenever $\mathcal{H}_t$ are of finite dimension or more generally of finite pseudo-dimension. Without restriction we can assume that $t = 0$. From (2.18) and (2.19) and the definition of the classes $\mathcal{A}_{\mathcal{H},t}$ it follows that

$$y_h(x) \equiv y_{h,0}(x) = \langle f(x), \mathbb{1}_{A_h}(x)\rangle, \tag{6.14}$$

where

$$\begin{aligned} \mathbb{1}_{A_h}(x) = \big(&1_{A_{h,0}}(x_0), 1_{A^c_{h,0}}(x_0)1_{A_{h,1}}(x_1), \ldots, \\ &1_{A^c_{h,0}}(x_0)\cdots 1_{A^c_{h,T-2}}(x_{T-2})1_{A_{h,T-1}}(x_{T-1}), 1_{A^c_{h,0}}(x_0)\cdots 1_{A^c_{h,T-1}}(x_{T-1})\big), \end{aligned} \tag{6.15}$$

and $\langle\cdot,\cdot\rangle$ is the ordinary euclidian scalar product in $\mathbb{R}^{T+1}$. Let $(x^1, \ldots, x^n) \subset \mathcal{R}$ be a fixed, but arbitrary subset of cardinality $n$ and let $(t_1, \ldots, t_n)$ be an arbitrary threshold vector. To bound the pseudo-dimension of $\mathcal{Y}_0$ on has to investigate the cardinality of

$$\begin{aligned} &\mathcal{S}(t_1, \ldots, t_n) = \\ &\qquad \{\big(\operatorname{sign}(\langle f(x^1), \mathbb{1}_{A_h}(x^1)\rangle - t_1), \ldots, \operatorname{sign}(\langle f(x^n), \mathbb{1}_{A_h}(x^n)\rangle - t_n)\big) \mid h \in \mathcal{H}\} \end{aligned}$$

as a subset of $\{0,1\}^n$. Let $\Delta_{\mathcal{C}}(m)$ be the growth function of a VC class, introduced in (5.19). As $h$ varies over $\mathcal{H}$ the first column of the matrix

$$\begin{pmatrix} \mathbb{1}_{A_h}(x^1) \\ \vdots \\ \mathbb{1}_{A_h}(x^n) \end{pmatrix}$$

can take on $\Delta_{\mathcal{A}_{\mathcal{H},0}}(n)$ different vectors in $\{0,1\}^n$ of the maximal possible cardinality $2^n$. For the subsequent columns corresponding to $t > 0$, additional degree of freedom is generated only by the indicator functions $1_{A_{h,t}}$. Consequently,

$$\#\mathcal{S}(t_1, \ldots, t_n) \leq \Delta_{\mathcal{A}_{\mathcal{H},0}}(n)\cdots\Delta_{\mathcal{A}_{\mathcal{H},T-1}}(n) \leq \left(\frac{e\,n}{k}\right)^{T\,k} < 2^n$$

for $n$ sufficiently large. Note that the above bound is independent of $(t_1, \ldots, t_n)$ and holds for any set of cardinality $n$. As a result, $\mathcal{Y}_0$ cannot pseudo-shatter sets of cardinality $n$ for $n$ large enough, because this would require at least $2^n$ different graphs over an $n$-point set. Finally, to bound the pseudo-dimension of $\mathcal{Y}_0$ observe that $\dim_P(\mathcal{Y}_0) \leq n_0$ if it cannot pseudo-shatter $n$-point set for $n > n_0$. By the above remarks, this is equivalent to

$$\#\mathcal{S}(t_1, \dots, t_n) < 2^n$$

for $n > n_0$ or by the above estimate,

$$\left(\frac{e\,n}{k}\right)^{T\,k} < 2^n$$

for $n > n_0$. Looking for solutions $n_0 = jk$ that are multiples of $k$ yields

$$T \log_2(ej) < j,$$

which is satisfied for example by $j = 2T \log_2(e\,T)$. □

**Corollary 6.3.** *Let $\mathcal{H} \subset L_2(X)$ be a finite dimensional linear approximation architecture. Then the classes $\mathcal{Y}_t, \mathcal{H}_t(r), \mathcal{L}_t(r)$ are Glivenko-Cantelli classes.*

*Proof.* All classes have an $L_1$ envelope. Because the $L_1$ metric is shorter than the $L_2$ metric hence, by (5.6) the above $L_2$ covering number bounds imply the required $L_1$ covering number bounds. □

The next step is to show that the sample minimizers $\hat{q}_{n,\mathcal{H}}$ remain almost surely in a compact set of the approximation architecture $\mathcal{H}$. It relies crucially on the convexity of the criterion function.

**Lemma 6.4** (Compactness Lemma)**.** *There exists a compact subset $\mathcal{K} \subset \mathcal{H}$ such that for all $h \in \mathcal{K}$, $q^*_{\mathcal{H},t}(h) \in \mathcal{H}_t \cap \mathcal{K}$ and $q^*_{n,\mathcal{H},t}(h) \in \mathcal{H}_t \cap \mathcal{K}$ almost surely for $n \to \infty$.*

*Proof.* Denote by $B_r(x)$ the metric ball of radius $r$ around $x$. The proof is by induction. For $t = T-1$, note that $q^*_{\mathcal{H},t}(h)$ and $q^*_{n,\mathcal{H},t}(h)$ are independent of $h$. Let $\varepsilon > 0$ arbitrary. Choose $r_{T-1} > 0$ large enough such that $B_\varepsilon(q^*_{\mathcal{H},t}(h)) \subset \mathcal{H}_{T-1}(r_{T-1})$. Because $\mathcal{L}_{T-1}(r_{T-1})$ is a Glivenko-Cantelli class,

$$\sup\nolimits_{u \in \mathcal{H}_{T-1}(r_{T-1}), h \in \mathcal{H}} |L_{n,T-1}(u,h) - L_{T-1}(u,h)| \to 0$$

almost sure. This and the convexity of $L_{T-1}(u,h)$ as a function of $u$ shows that $L_{n,T-1}(u,h)$ must have a local minimum in $B_\varepsilon(q^*_{\mathcal{H},t}(h))$ almost surely for $n$ large enough. By the convexity of $L_{n,T-1}(u,h)$, this local minimum must be the unique global minimum $q^*_{n,\mathcal{H},t}(h)$ of $L_{n,T-1}(u,h)$.

Now let $t < T-1$. Choose $r_t > 0$ large enough such that $B_\varepsilon(q^*_{\mathcal{H},t}(h)) \subset \mathcal{H}_t(r_t)$ for all $h \in \mathcal{H}$ with $h_s \in \mathcal{H}_s(r_s)$, $s > t$. Fix an $h$ with this property. The same argument as before shows that $q^*_{n,\mathcal{H},t}(h) \in B_\varepsilon(q^*_{\mathcal{H},t}(h))$ almost surely for $n$ large enough. □

**Proposition 6.5** (Convergence of Parameterized Minimizers)**.**

$$\lim\nolimits_{n\to\infty} \|q^*_{n,\mathcal{H},t}(h)(X_t) - q^*_{\mathcal{H},t}(h)(X_t)\|_{2,P} = 0, \tag{6.16}$$

*$\mathbb{P}$-almost surely, uniformly in $h \in \mathcal{K}$.*

*Proof.* From (6.10) and (6.8) it follows that for an arbitrary $h \in \mathcal{K}$

$$\begin{aligned} P\big(q^*_{n,\mathcal{H},t}(h) - q^*_{\mathcal{H},t}(h)\big)^2(X) = L_t(\hat{q}_n, h) - L_t(q^*, h) \leq \\ 2\sup\nolimits_{u \in \mathcal{H}_t \cap \mathcal{K}} |L_{t,n}(u,h) - L_t(u,h)| \leq \\ 2\sup\nolimits_{u \in \mathcal{H}_t \cap \mathcal{K}, h \in \mathcal{K}} |L_{t,n}(u,h) - L_t(u,h)|. \end{aligned} \tag{6.17}$$

By compactness, $\mathcal{H}_t \cap \mathcal{K} \times \mathcal{K}$ has an envelope in $L_2$ and is therefore a Glivenko-Cantelli class. The right hand side of the above estimate converges to zero almost surely. □

**Lemma 6.6** (Continuity Lemma). *Assume that $f \in L_2(X)$ and that $h \in L_2(X)$ satisfies*

$$P\big(h(X)_t = f(X)_t\big) = 0, \quad \forall\, t. \tag{6.18}$$

*Let $g_n : \Omega^\infty \to L_2(X)$ a sequence of random elements, such that for all $s \geq t$, $g_n(X)_s \to h(X)_s$ in probability, $\mathbb{P}$-almost sure for $n \to \infty$, that is for every $\varepsilon > 0$,*

$$\lim\nolimits_{n\to\infty} P\big(|g_n(X)_s - h(X)_s| > \varepsilon\big) = 0, \tag{6.19}$$

*$\mathbb{P}$-almost sure. Then*

$$\lim\nolimits_{n\to\infty} \|y_{g_n,t}(X) - y_{h,t}(X)\|_{2,P} = 0 \tag{6.20}$$

*$\mathbb{P}$-almost sure.*

*Proof.* Integrate the point-wise estimate (2.23).

$$\|y_{g_n,t}(X) - y_{h,t}(X)\|_{2,P} \leq \sum_{s=t}^{T-1} \sum_{r=s}^{T-1} \|f(X)_r\, 1_{\{|f(X)_s - h(X)_s| \leq |g_n(X)_s - h(X)_s|\}}\|_{2,P}. \tag{6.21}$$

To proceed further, let $a > 0$ be arbitrary. Then,

$$P\big(f^2(X)_r\, 1_{\{|f(X)_s-h(X)_s|\leq|g_n(X)_s-h(X)_s|\}}\big) \leq aP\big(|f(X)_s - h(X)_s| \leq |g_n(X)_s - h(X)_s|\big) + P\big(f^2(X)_r\, 1_{\{f^2(X)_r > a\}}\big). \tag{6.22}$$

Let $\eta > 0$. Choose $a$ large enough such that the last term in (6.22) is smaller than $\eta$. Now for $\varepsilon > 0$ arbitrary

$$aP\big(|f(X)_s - h(X)_s| \leq |g_n(X)_s - h(X)_s|\big) \leq aP\big(|f(X)_s - h(X)_s| \leq \varepsilon\big) + aP\big(|g_n(X)_s - h(X)_s| > \varepsilon\big). \tag{6.23}$$

By (6.18), one can choose $\varepsilon > 0$ small enough such that the first term is less than $\eta$ and then select $n_0$ large enough such the last term in (6.23) is bounded by $\eta$ for all $n > n_0$, $\mathbb{P}$-almost sure. Consequently,

$$P\big(f^2(X)_r\, 1_{\{|f(X)_s-h(X)_s|\leq|g_n(X)_s-h(X)_s|\}}\big) \leq 3\eta$$

for $n > n_0$, $\mathbb{P}$-almost sure. □

*Remark* 6.7. The conclusion of lemma 6.6 holds in particular if

$$\|g_n(X)_s - h(X)_s\|_{2,P} \to 0,$$

$\mathbb{P}$-almost sure.

**Theorem 6.8.** *Assume that the payoff $f$ is in $L_2(X)$ and that*

$$P\big(q_{\mathcal{H}}(X)_t = f(X)_t\big) = 0, \quad \forall\, t. \tag{6.24}$$

*Then, the sequence of random elements $\hat{q}_{n,\mathcal{H}} : \Omega^\infty \to \mathcal{H} \subset L_2(X)$ converges $\mathbb{P}$-almost surely to $q_{\mathcal{H}} \in \mathcal{H}$ in the norm of $L_2(X)$ for $n \to \infty$.*

*Proof.* The proof follows by induction. Noting that $q^*_{\mathcal{H},T-1}(h)$ is constant in $h$, the case of $t = T-1$ is already established by (6.16). Let $t < T-1$ and make use of the recursions $q_{\mathcal{H},t} = q^*_{\mathcal{H},t}(q_{\mathcal{H}})$ and $\hat{q}_{n,\mathcal{H},t} = q^*_{n,\mathcal{H},t}(\hat{q}_{n,\mathcal{H}})$. Then

$$\|q_{\mathcal{H}}(X)_t - \hat{q}_{n,\mathcal{H}}(X)_t\|_{2,P} \leq \|q^*_{\mathcal{H},t}(q_{\mathcal{H}})(X_t) - q^*_{\mathcal{H},t}(\hat{q}_{n,\mathcal{H}})(X_t)\|_{2,P} + \|q^*_{\mathcal{H},t}(\hat{q}_{n,\mathcal{H}})(X_t) - q^*_{n,\mathcal{H},t}(\hat{q}_{n,\mathcal{H}})(X_t)\|_{2,P}.$$

For the first term,

$$\|q^*_{\mathcal{H},t}(q_{\mathcal{H}})(X_t) - q^*_{\mathcal{H},t}(\hat{q}_{n,\mathcal{H}})(X_t)\|_{2,P} = \|\mathrm{pr}_{\mathcal{H}_t}\big(y_{q_{\mathcal{H}},t+1}(X) - y_{\hat{q}_{n,\mathcal{H}},t+1}(X)\big)\|_{2,P}.$$

By lemma 6.6 and the induction hypothesis, this converges to zero almost surely as well. The second term converges to zero almost surely by lemma 6.4 and proposition 6.5. □

6.2. **Convergence of Stopping Times.** In section 2.8 it has been noted that the functionals $\tau_{q,t}$ are not continuous on all of $\mathcal{R}$. However, conditions like

$$P\big(q_{\mathcal{H}}(X)_t = f(X)_t\big) = P\big(q(X)_t = f(X)_t\big) = 0, \quad \forall\, t. \tag{6.25}$$

enforce that the set of discontinuities have $P$-measure zero, hence weak convergence of $\tau_{q_{\mathcal{H}},t}(X)$ to $\tau_{q,t}(X)$ for $\dim \mathcal{H} \to \infty$ follows from the generalized continuous mapping theorem. Proposition 2.4 allows a refined analysis.

**Theorem 6.9.** *Assume that $f \in L_2(X)$ and that (6.25) holds for a sequence of approximation architectures $\mathcal{H}^k$ of dimension $k$. If $q_{\mathcal{H}^k} \to q$ in probability for $k \to \infty$, then*

$$\lim_{k\to\infty} \|\tau_t^* - \tau_{q_{\mathcal{H}^k},t}(X)\|_{2,P} = 0. \tag{6.26}$$

*If $\hat{q}_{n,\mathcal{H}^k} \to q_{\mathcal{H}^k}$ in probability for $k \to \infty$, $\mathbb{P}$-almost sure, that is for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\big(|\hat{q}_{n,\mathcal{H}^k}(X)_s - q_{\mathcal{H}^k}(X)_s| > \varepsilon\big) = 0$$

*$\mathbb{P}$-almost sure, then for every $k$*

$$\lim_{n\to\infty} \|\tau_{q_{\mathcal{H}^k},t}(X) - \tau_{\hat{q}_{n,\mathcal{H}^k},t}(X)\|_{2,P} = 0. \tag{6.27}$$

*$\mathbb{P}$-almost sure.*

*Proof.* Note that $\tau_t^* = \tau_{q,t}(X)$. From proposition 2.4

$$\|\tau_{q,t}(X) - \tau_{q_{\mathcal{H}^k},t}(X)\|_{2,P} \leq \sum_{s=t}^{T-1} \big(s + \ldots + T\big)\, P\big(|f(X)_s - q(X)_s| \leq |q(X)_s - q_{\mathcal{H}^k}(X)_s|\big).$$

But for $\varepsilon > 0$

$$P\big(|f(X)_s - q(X)_s| \leq |q(X)_s - q_{\mathcal{H}^k}(X)_s|\big) \leq P\big(|f(X)_s - q(X)_s| \leq \varepsilon\big) + P\big(|q(X)_s - q_{\mathcal{H}^k}(X)_s| > \varepsilon\big),$$

which can be made arbitrarily small for $\varepsilon$ small enough and $k$ large enough, (6.26) follows. The reasoning for (6.27) is similar as the above estimates hold $\mathbb{P}$-almost sure for $q_{\mathcal{H}^k}$ in place of $q$ and $\hat{q}_{n,\mathcal{H}^k}$ in place of $q_{\mathcal{H}^k}$. □

6.3. **Error Probabilities and Sample Complexity.** Let $\mathcal{H}$ be a linear approximation architecture of dimension $k$. The sample complexity function $n(k, \varepsilon, \eta)$ is defined as

$$n(k, \varepsilon, \eta) = \inf\{n \mid \mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}} - q_{\mathcal{H}}\|_{2,X}^2 > \varepsilon\big) \leq \eta\}. \tag{6.28}$$

It provides a worst case measure for the number of samples required to achieve a small sampling error with high confidence.

To efficiently estimate the error probabilities $\mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}} - q_{\mathcal{H}}\|_{2,X}^2 > \varepsilon\big)$ in terms of exponential inequalities we impose the following stronger assumption on the payoff function and the approximation architecture.

**Hypothesis 6.10.** *Assume $f \in L_\infty(X)$, that $\mathcal{H}$ is a finite dimensional linear approximation architecture satisfying $\mathcal{H} \subset L_\infty(X)$, and that for all $t$,*

$$P\big(|q_{\mathcal{H}}(X)_t - f(X)_t| \leq x\big) = o(x) \tag{6.29}$$

*as $x \to 0$.*

Note that in view of proposition 2.1, restricting to bounded payoff functions is feasible. It can be verified that condition (6.29) holds if the random variable $|q_{\mathcal{H}}(X)_t - f(X)_t|$ has a bounded density near 0. This is the case for the relevant practical examples. However, a stronger decay rate is typically false as can be seen even in the simplest case of an American put option and a polynomial approximation architecture.

Under hypothesis 6.10 there exists $r > 0$ such that $l_t(u,h) \in \mathcal{L}_t(r)$ for $u \in \mathcal{H}_t \cap \mathcal{K}, h \in \mathcal{K}$ and $\mathcal{L}_t(r)$ is uniformly bounded by some constant $L = (H+Y)^2$, where $Y = \|f\|_{\infty,X}$ and $H$ is an uniform bound for $\mathcal{H}_t(r)$. Note that the bound $H$ is not explicit. We therefore may want to restrict the approximation architecture explicitly to a compact domain. This compact domain is required to grow with increasing sample size. This leads to sieve estimators, see section 8.

The following recursive error probability estimate can now be proved.

**Proposition 6.11.** *Impose hypothesis 6.10. Then there exists a constant $C(\mathcal{H})$, depending on $\mathcal{H}$, such that for all $\varepsilon > 0$*

$$\begin{aligned}\mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}}(X)_t - q_{\mathcal{H}}(X)_t\|_{2,P}^2 > \varepsilon\big) \leq \\ \sum_{s=t+1}^{T-1} \mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s\|_{2,P}^2 > \frac{\varepsilon^2}{C(\mathcal{H})\|f\|_{\infty,X}^4}\big) + \\ \mathbb{P}\big(\sup_{u\in\mathcal{H}_t\cap\mathcal{K},h\in\mathcal{K}}|L_{t,n}(u,h) - L_t(u,h)| > \frac{\varepsilon}{2}\big).\end{aligned} \tag{6.30}$$

*Proof.* As in the proof of theorem 6.8

$$\begin{aligned}\mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}}(X)_t - q_{\mathcal{H}}(X)_t\|_{2,P} > \sqrt{\varepsilon}\big) \leq \\ \mathbb{P}\big(\|q^*_{\mathcal{H},t}(q_{\mathcal{H}})(X)_t - q^*_{\mathcal{H},t}(\hat{q}_{n,\mathcal{H}})(X)_t\|_{2,P} > \sqrt{\varepsilon}\big) + \\ \mathbb{P}\big(\|q^*_{\mathcal{H},t}(\hat{q}_{n,\mathcal{H}})(X)_t - q^*_{n,\mathcal{H},t}(\hat{q}_{n,\mathcal{H}})(X)_t\|_{2,P} > \sqrt{\varepsilon}\big) = I_1 + I_2\end{aligned}$$

For the first term by the definition of $q^*_{\mathcal{H},t}$, proposition 2.4 and the triangle inequality

$$\begin{aligned}I_1 \leq \mathbb{P}\big(\|y_{q_{\mathcal{H}},t+1}(X)_t - y_{\hat{q}_{n,\mathcal{H}},t+1}(X)_t\|_{2,P} > \sqrt{\varepsilon}\big) \leq \\ \sum_{s=t+1}^{T-1} \mathbb{P}\big(\|f\|_{\infty,X} P(|f(X)_s - q_{\mathcal{H}}(X)_s| \leq |\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s|)^{1/2} > \sqrt{\varepsilon}\big)\end{aligned}$$

But

$$P\big(|f(X)_s - q_{\mathcal{H}}(X)_s| \leq |\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s|\big) \leq c\big(\|\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s\|_{\infty,P}\big). \tag{6.31}$$

Therefore

$$I_1 \leq \sum_{s=t+1}^{T-1} \mathbb{P}\Big(\|\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s\|_{\infty,P} > \frac{\varepsilon}{\|f\|_{\infty,X}^2}\Big)$$

Now on a finite dimensional vector space all norms are equivalent. The bound for $I_2$ follows directly from (6.17). □

*Remark* 6.12. For specific linear approximation architectures $\mathcal{H}$ the $L_2$-norm and the $L_\infty$ norm are connected by an inequality of the form

$$\|h(X)_t\|_{\infty,P} \leq c\sqrt{k}\|h(X)_t\|_{2,P} \tag{6.32}$$

for all $h \in \mathcal{H}_t$, where $k = \dim(\mathcal{H}_t)$. Examples are architectures generated by uniformly bounded basis functions such as trigonometric systems, piecewise polynomials, splines but also wavelets. This makes the constant $C(\mathcal{H})$ appearing in more explicit.

It is important to note that in the first term on the right hand side of (6.30) deteriorates the estimate by an additional power in $\varepsilon$. The second term can be bounded by the uniform deviation inequality (5.33) of Talagrand.

**Proposition 6.13.**

$$\mathbb{P}\big(\sup_{l\in\mathcal{L}_t(r)}|P_nl-Pl|>\varepsilon\big)\leq\left(C(V)\frac{\sqrt{n}\,\varepsilon}{\sqrt{D}\,L}\right)^D\exp\left(-\frac{2n\varepsilon^2}{L^2}\right)\tag{6.33}$$

*where*

$$D=4\big(d_{\mathcal{H}_t}+d_{\mathcal{Y}_t}\big),\quad V=\big(e^2(d_{\mathcal{H}_t}+1)(d_{\mathcal{Y}_t}+1)\big)^{1/D}\big(2^{13}e\big)^{1/4}.$$

*Proof.* From lemma (5.3) and theorem 5.6

$$N(\varepsilon L,\mathcal{L}_t(r),d_{2,P})\leq N(\frac{\varepsilon}{2}(H+Y),\mathcal{H}_t-\mathcal{Y}_t,d_{4,P})\leq$$
$$N(\frac{\varepsilon}{4}(H+Y),\mathcal{H}_t(r),d_{4,P})N(\frac{\varepsilon}{4}(H+Y),\mathcal{Y}_t,d_{4,P})\leq$$
$$N(\frac{\varepsilon}{4}H,\mathcal{H}_t(r),d_{4,P})N(\frac{\varepsilon}{4}Y,\mathcal{Y}_t,d_{4,P})\leq e^2(d_{\mathcal{H}_t}+1)(d_{\mathcal{Y}_t}+1)\left(\frac{2^{13}e}{\varepsilon^4}\right)^{d_{\mathcal{H}_t}+d_{\mathcal{Y}_t}}$$

Then apply (5.33). □

It is possible to lower $d$ to $2(d_{\mathcal{H}_t}+d_{\mathcal{Y}_t})$ by using $d_{2,P}$ in place of $d_{4,P}$. However, in this case the constant $V$ would depend on $\|f\|_{\infty,X}$ and on the bound $L$.

*Remark* 6.14. Vapnik [45, 46] provides uniform onesided deviation inequalities as well. However, they require that the classes of functions are so called VC major classes. He also proved deviation inequalities for classes of nonnegative functions satisfying some moment inequalities. We only showed that the classes $\mathcal{L}$ are VC subgraph classes. This is why we prefer to apply Talagrand's estimate. For basic relation between the two concepts see [13] and [42].

We can now combine propositions 6.12 and 6.13 to obtain an exponential error probability estimate.

**Theorem 6.15.** *Assume hypothesis 6.10 holds. Then*

$$\mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}}(X)_t-q_{\mathcal{H}}(X)_t\|_{2,P}^2>\varepsilon\big)\leq$$
$$\leq(T-t)2^{T-t}\left(C(V)\frac{\sqrt{n}\,\varepsilon}{\sqrt{D}\,L}\right)^D\exp\left(-\frac{n}{2L^2}\Big(\frac{\varepsilon}{C(\mathcal{H})\|f\|_{\infty,P}^4}\Big)^{2^{T-t-1}}\right)\tag{6.34}$$

*where $D$ and $V$ are as in proposition 6.13 and $C(\mathcal{H})$ is from proposition 6.12.*

*Proof.* Set $a_t(n)=\|\hat{q}_{n,\mathcal{H}}(X)_t-q_{\mathcal{H}}(X)_t\|_{2,P}^2$ and $b_t(n)=\sup_{l\in\mathcal{L}_t(r)}|P_nl-Pl|$. From proposition 6.12

$$\mathbb{P}\big(a_t(n)>\varepsilon\big)\leq\mathbb{P}\big(b_t(n)>\frac{\varepsilon}{2}\big)+\mathbb{P}\big(b_{t+1}(n)>\frac{\varepsilon_{t+1}}{2}\big)+$$
$$\sum_{s=t+2}^{T-1}\Big(\mathbb{P}\big(a_s(n)>\varepsilon_{t+1}\big)+\mathbb{P}\big(a_s(n)>\frac{\varepsilon_{t+1}^2}{C(\mathcal{H})\|f\|_{\infty,P}^4}\big)\Big),$$

where $\varepsilon_t=\varepsilon$ and $\varepsilon_{s+1}=\varepsilon_s^2/(C(\mathcal{H})\|f\|_{\infty,P}^4)$ is defined recursively for $s>t$. Without restriction we may assume that $C(\mathcal{H})\|f\|_{\infty,X}^4\geq 1$, hence

$$\mathbb{P}\big(a_t(n)>\varepsilon\big)\leq\mathbb{P}\big(b_t(n)>\frac{\varepsilon}{2}\big)+\mathbb{P}\big(b_{t+1}(n)>\frac{\varepsilon_{t+1}}{2}\big)+$$
$$2\sum_{s=t+2}^{T-1}\mathbb{P}\big(a_s(n)>\varepsilon_{t+2}\big)\leq$$
$$\mathbb{P}\big(b_t(n)>\frac{\varepsilon}{2}\big)+\sum_{s=t+1}^{T-1}2^{s-t-1}\mathbb{P}\big(b_s(n)>\frac{\varepsilon_s}{2}\big),$$

and from proposition 6.13

$$\mathbb{P}\big(a_t(n) > \varepsilon\big) \leq (T-t)2^{T-t}\left(C(V)\frac{\sqrt{n}\,\varepsilon}{\sqrt{d}\,L}\right)^d \exp\Big(-\frac{n\varepsilon_{T-1}^2}{2L^2}\Big).$$

Now apply the defining recursion for $\varepsilon_s$ to derive the final estimate (6.34). □

Consequently, increasing the number of time steps causes arbitrarily slow exponential convergence of the error probability. This unpleasant feature of the Longstaff-Schwartz algorithm is primarily a consequence of the lack of smoothness of the functional $y_{t,h}$ in $h$, which leads to a non-negligible error propagation in the backward recursion. A potential improvement is sketched in section 9.

## 7. Central Limit Theorem

The advantage of the empirical process setup is that Central Limit Theorems can be proved by a delta method, which builds on a first order Taylor expansion of the criterion function. It turns out that the regularity condition required for the remainder term is a sort of stochastic equicontinuity condition, which is typically verified by empirical process techniques. The following result for sequences that come close enough to sample minimizers is particularly convenient for our application.

**Theorem 7.1** (Pollard, [35, Section VII.1, Theorem 5])**.** *Let $(\Omega, P)$ be a probability space, $V$ a finite dimensional vector space with scalar product $\langle\cdot,\cdot\rangle$, and $\mathcal{K} \subset V$ a parameter set. Let $l : \mathcal{K}\times\Omega \to \mathbb{R}$ and define $L(h) = Pl(\cdot, h)$, $L_n(h) = P_n l(\cdot, h)$. Introduce the minimizers by $h^* = argmin_{h\in\mathcal{K}} L(h)$ and $h_n^* = argmin_{h\in\mathcal{H}} L_n(h)$. Assume that the function $l$ satisfies in a neighborhood of $h^*$ the expansion*

$$l(h,\cdot) = l(h^*,\cdot) + \langle\Delta_{h^*}(\cdot), h-h^*\rangle + \|h-h^*\|\, r_{h^*}(h,\cdot) \tag{7.1}$$

*for some functions $\Delta = \Delta_{h^*} : \Omega \to V$ and $r_{h^*}(h,\cdot) : \Omega \to \mathbb{R}$. Let $h_n \in \mathcal{K}$ be a sequence such that*

$$L_n(h_n) = \inf_{h\in\mathcal{K}} L_n(h) + o_P(n^{-1}). \tag{7.2}$$

*Assume furthermore*

(i) *$h^*$ is an interior point of $\mathcal{K}$,*
(ii) *$L$ has a non-singular $2^{nd}$ derivative matrix $\Gamma$ at $h^*$.*
(iii) *$\Delta_{h^*}(\cdot) \in \mathcal{L}_2(\Omega, P)$,*
(iv) *$E_n$ indexed by $\{r_{h^*}(h,\cdot) \mid h \in \mathcal{K}\}$ is stochastically equicontinuous at $h^*$.*

*Then, the Central Limit Theorem holds, i.e.,*

$$\sqrt{n}\big(h_n - h^*\big) \xrightarrow{w} N\big(0, \Gamma^{-1}(P(\Delta\Delta^t) - P(\Delta)P(\Delta)^t)\Gamma^{-1}\big). \tag{7.3}$$

Arcones investigated in a series of papers approximate $M$-estimators and their limit behavior. In particular he determined the rate for (7.2), such that the approximate $M$-estimator is asymptotically normal with rate $n^{1/2}$. We refer to [1] for more details.

**Corollary 7.2.** *Let $\|\cdot\|$ be a norm on $V$. Then*

$$\|h_n - h^*\| = O_P(n^{-1/2}). \tag{7.4}$$

*Proof.* Abbreviate $\Gamma^{-1}(P(\Delta\Delta^t) - P(\Delta)P(\Delta)^t)\Gamma^{-1}$ by $G$. Let $\varepsilon > 0$ and take $M_\varepsilon$ such that $P\big(\|N(0,G)\| > M_\varepsilon\big) < \varepsilon$. But by weak convergence (7.3)

$$|P\big(n^{1/2}\|h_n - h^*\| > M_\varepsilon\big) - P\big(\|N(0,G)\| > M_\varepsilon\big)| < \varepsilon,$$

for $n$ sufficiently large, hence

$$P\big(n^{1/2}\|h_n - h^*\| > M_\varepsilon\big) < 2\varepsilon,$$

which proves (7.4) by definition of $O_P$. □

For every fixed $t$, theorem 7.1 applies to the Longstaff-Schwartz algorithm via the translation

$$l(h, \omega) = l_t(h, q_{\mathcal{H}})(X(\omega)),$$

and $L(h) = L_t(h, q_{\mathcal{H}})$, $L_n(h) = L_{t,n}(h, q_{\mathcal{H}})$. Consequently, the roles of $h_n^*$ and $h_n$ are taken by

$$\begin{aligned} h_n^* &= \bar{q}_{n,\mathcal{H},t} = q^*_{n,\mathcal{H},t}(q_{\mathcal{H}}), \\ h_n &= \hat{q}_{n,\mathcal{H},t} = q^*_{n,\mathcal{H},t}(\hat{q}_{n,\mathcal{H}}). \end{aligned} \tag{7.5}$$

Expand every $g \in \mathcal{H}_t$ in a basis $b_t^i$ and identify $g = \sum_{i=1}^k \lambda_i(g)\, b_t^i$ with $\lambda(g) \in \mathbb{R}^k$. Then (6.2) can be expressed as

$$l_t(g, u) = l_t(h, u) + \langle \Delta_{h,u}(x), \lambda(g-h)\rangle + \frac{1}{2}\langle \Gamma_{h,u}(x)\lambda(g-h), \lambda(g-h)\rangle, \tag{7.6}$$

where

$$\Delta_{h,u} : \mathcal{R} \to \mathbb{R}^k, \quad \Delta_{h,u}(x) = 2\big(h(x_t) - y_{u,t+1}(x)\big) \begin{pmatrix} b_1(x_t) \\ \vdots \\ b_k(x_t) \end{pmatrix} \tag{7.7}$$

and

$$\Gamma_{h,u} : \mathcal{R} \to \mathbb{R}^{k\times k}, \quad \Gamma_{h,u}(x)_{i,j} = 2 b_i(x_t) b_j(x_t) \tag{7.8}$$

corresponds to the gradient respectively Hessian and $\langle\cdot,\cdot\rangle$ is the euclidian scalar product. This establishes 7.1. Assumptions (i)-(iii) can be checked in a straightforward manner, whereas (iv) can be verified by empirical process techniques. Condition (7.2) is more involved and depends on the continuity of the functional $q^*_{n,\mathcal{H},t}$.

**Proposition 7.3.** *Assume that in addition to 6.10,*

$$\|\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s\|_{2,P} = O_{\mathbb{P}}(n^{-1/2}) \tag{7.9}$$

*for $s > t$, then, condition (7.2) is satisfied for $t$, i.e., for every $\varepsilon > 0$,*

$$\mathbb{P}\big(|L_{t,n}(\bar{q}_{n,\mathcal{H},t}, q_{\mathcal{H}}) - L_{t,n}(\hat{q}_{n,\mathcal{H},t}, q_{\mathcal{H}})| \geq \frac{\varepsilon}{n}\big) \to 0 \tag{7.10}$$

*for $n \to \infty$.*

*Proof.* Integrating (6.2) with respect to $P_n$ gives

$$L_{n,t}(\hat{q}_{n,\mathcal{H},t}, q_{\mathcal{H}}) - L_{n,t}(\bar{q}_{n,\mathcal{H},t}, q_{\mathcal{H}}) = P_n\big(q^*_{n,\mathcal{H},t}(\hat{q}_{n,\mathcal{H}})(X_t) - q^*_{n,\mathcal{H},t}(q_{\mathcal{H}})(X_t)\big)^2.$$

By the continuity lemma 6.6, if $\hat{q}_{n,\mathcal{H},s}$ is close to $q_{\mathcal{H},s}$, for $s > t$, the right hand side is expected to be small as well. To make this precise, note that the sample minimizers $q^*_{n,\mathcal{H},t}(h)$ are characterized by

$$D_1 L_{n,t}(q^*_{n,\mathcal{H},t}(h), h) = 0,$$

which is equivalent to

$$P_n\big((q^*_{n,\mathcal{H},t}(h)(X_t) - y_{h,t+1}(X))\, v_t(X_t)\big) = 0 \quad \forall\, v_t \in \mathcal{H}_t.$$

Apply this to obtain

$$P_n\big((\hat{q}_{n,\mathcal{H},t} - \bar{q}_{n,\mathcal{H},t})(X_t)\, v_t(X_t)\big) = P_n\big((y_{q_{\mathcal{H}},t+1} - y_{\hat{q}_{n,\mathcal{H}},t+1})(X)\, v_t(X_t)\big).$$

Set $v_t = \hat{q}_{n,\mathcal{H},t} - \bar{q}_{n,\mathcal{H},t} \in \mathcal{H}_t$ and apply the Cauchy-Schwarz inequality to get

$$\|(\hat{q}_{n,\mathcal{H},t} - \bar{q}_{n,\mathcal{H},t})(X_t)\|_{2,P_n} \leq \|(y_{q_{\mathcal{H}},t+1} - y_{\hat{q}_{n,\mathcal{H}},t+1})(X)\|_{2,P_n}.$$

The class $\mathcal{Y}_t$ is Glivenko-Cantelli. Therefore

$$\|(y_{q_{\mathcal{H}},t+1} - y_{\hat{q}_{n,\mathcal{H}},t+1})(X)\|_{2,P_n} \leq 2\|(y_{q_{\mathcal{H}},t+1} - y_{\hat{q}_{n,\mathcal{H}},t+1})(X)\|_{2,P}$$

almost sure, as $n \to \infty$. The point-wise estimate (2.23) of proposition 2.4 yields

$$\|(y_{q_{\mathcal{H}},t+1} - y_{\hat{q}_{n,\mathcal{H}},t+1})(X)\|_{2,P} \leq \\ \|f\|_{\infty,X} \sum_{s=t+1}^{T-1} (T-s)\, P\big(|f(X)_s - q_{\mathcal{H}}(X)_s| \leq |\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s|\big).$$

Note that on a finite dimensional vector space all norms are equivalent. Consequently, if (7.9) holds for $\|\cdot\|_{2,P}$ it also holds for $\|\cdot\|_{\infty,P}$ and hypotheses 6.10 imply

$$P\big(|f(X)_s - q_{\mathcal{H}}(X)_s| \leq |\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s|\big) \leq \\ P\big(|f(X)_s - q_{\mathcal{H}}(X)_s| \leq \|\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s\|_{\infty,P}\big) \leq \\ o\big(\|\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s\|_{\infty,P}\big) \leq o_{\mathbb{P}}(n^{-1/2}),$$

hence

$$|L_{n,t}(\hat{q}_{n,\mathcal{H},t}, q_{\mathcal{H}}) - L_{n,t}(\bar{q}_{n,\mathcal{H},t}, q_{\mathcal{H}})| \leq \|(\hat{q}_{n,\mathcal{H},t} - \bar{q}_{n,\mathcal{H},t})(X_t)\|^2_{2,P_n} = o_{\mathbb{P}}(n^{-1}),$$

proving (7.10). □

**Proposition 7.4.** *Condition (i)-(iv) of theorem 7.1 are satisfied.*

*Proof.* Condition (i), (ii) are obviously satisfied by lemma 6.4 choosing $\mathcal{K}$ sufficiently large. (iii) holds whenever $f \in L_2(X)$. As for condition (iv)

$$\mathcal{R} = \Big\{ r_{q_{\mathcal{H},t}}(h) = \frac{\langle \Gamma_{q_{\mathcal{H},t},q_{\mathcal{H}}}(x)\lambda(h - q_{\mathcal{H}}), \lambda(h - q_{\mathcal{H}})\rangle}{2\|\lambda(h - q_{\mathcal{H}})\|} \mid h \in \mathcal{H}_t \Big\} \tag{7.11}$$

$\mathcal{R}$ is a subset of the linear finite dimensional space spanned by $\{1/2\, b_t^i\, b_t^j \mid 1 \leq i \leq j \leq k\}$. As such it is a VC subgraph class. The coefficients of $r \in \mathcal{R}$ are restricted to euclidian length 1. Therefore, $\mathcal{R}$ has an envelope in $L_\infty(X)$, hence in $L_2(X)$ and is a Donsker class by corollary 5.13. Condition (iv) is satisfies by theorem 5.11. □

We can now derive a Central Limit Theorem for the Longstaff-Schwartz algorithm.

**Theorem 7.5.** *Assume that hypothesis 6.10 holds. Select for every approximation space $\mathcal{H}_t$ a basis $b_t^i$ and set*

$$\Delta = \Delta_{q_{\mathcal{H},t},q_{\mathcal{H}}}, \quad \Gamma = \Gamma_{q_{\mathcal{H},t},q_{\mathcal{H}}}. \tag{7.12}$$

*Then*

$$\sqrt{n}\big(\hat{q}_{n,\mathcal{H},t} - q_{\mathcal{H},t}\big) \xrightarrow{w} N\big(0, \Gamma^{-1}(P(\Delta\Delta^t) - P(\Delta)P(\Delta)^t)\Gamma^{-1}\big). \tag{7.13}$$

*Proof.* The proof is by induction. For $t = T-1$, condition (7.2) is void because, $q^*_{n,\mathcal{H},T-1}(h)$ is constant in $h$, hence $\bar{q}_{n,\mathcal{H},T-1} = \hat{q}_{n,\mathcal{H},T-1}$. Theorem 7.1 applies and yields the desired weak convergence. As for $t < T-1$, by corollary 7.2, (7.9) holds for all $s > t$ and proposition 7.3 implies (7.2). Theorem 7.1 applies again. □

## 8. A Sieved Longstaff-Schwartz Algorithm

A Uniform Law of Large Numbers may or may not hold, depending on the size of the approximation architecture $\mathcal{H}$, the structure of the loss functions $l_t$, and on the sampling distribution $P$. If the class of loss functions is too large, empirical risk minimization may fail to converge. A rescue in such a case is a refinement of empirical risk minimization called structural risk minimization or sieve estimation, [46, 42]. Sieve estimation performs empirical risk minimization on a sequence of approximation architectures $\mathcal{H}^{(n)}$, called sieves, designed to grow with increasing sample size in a suitable manner such that good approximation properties as well as stable estimators can be guaranteed. This approach leads to a whole new family of algorithms for solving optimal stopping problems.

We provide one example of such a sieve estimator. Assume that $\mathbf{H}^{(n)}$ is a sequence of linear approximation architectures of dimension $k_n$ growing dense in $L_2(X)$ for $k_n \to \infty$. Let $H_n > 0$ and define the sieves

$$\mathcal{H}_t^{(n)} = \mathbf{H}_t^{(n)}(H_n) \tag{8.1}$$

and define the classes $\mathcal{Y}_t^{(n)}$ and $\mathcal{L}_t^{(n)}$ as in (6.11) but with the compact approximation architecture $\mathcal{H}^{(n)}$. This amounts of truncating the coefficients relative to a basis. In particular it enforces the compactness lemma 6.4 by construction of the algorithm. This is also of importance for more general approximation architectures.

How do $k_n$ and $H_n$ need to converge to $\infty$ such that we have convergence in probability? Before we can proof such a consistency result, we need a version of proposition 6.12 which does not depend on the approximation architecture.

**Hypothesis 8.1.** *Let $\mathcal{H}^{(n)} = \mathbf{H}_t^{(n)}(H_n)$ be a sequence of approximation architectures, where $\mathbf{H}^{(n)} \subset L_\infty(X)$ is a linear subspace of dimension $k_n$. Assume that $\cup_n \mathcal{H}^{(n)}$ is dense in $L_2(X)$ and that*

$$P\big(|q_{\mathcal{H}^{(n)}}(X)_t - f(X)_t| \leq x\big) \leq Cx \tag{8.2}$$

*as $x \to 0$, for a constant $C$ independent of $n$.*

**Proposition 8.2.** *Let $f \in L_\infty(X)$ and impose hypothesis 8.1. Then for all $\varepsilon > 0$*

$$\begin{aligned}\mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}^{(n)}}(X)_t - q_{\mathcal{H}^{(n)}}(X)_t\|_{2,P}^2 > \varepsilon\big) \leq \\ \sum_{s=t+1}^{T-1} \mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}^{(n)}}(X)_s - q_{\mathcal{H}^{(n)}}(X)_s\|_{2,P}^2 > \frac{4\varepsilon^3}{27C^2\|f\|_{\infty,X}^6}\big) + \\ \mathbb{P}\big(\sup\nolimits_{l \in \mathcal{L}_t^{(n)}} |P_n l - Pl| > \frac{\varepsilon}{2}\big).\end{aligned} \tag{8.3}$$

*Proof.* Fix $n$ and let $\mathcal{H} = \mathcal{H}^{(n)}$. Start as in the proof of 6.12. Instead of apply now the following estimate. For all $\eta > 0$,

$$\begin{aligned}P\big(|f(X)_s - q_{\mathcal{H}}(X)_s| \leq |\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s|\big) \leq \\ P\big(|f(X)_s - q_{\mathcal{H}}(X)_s| \leq \eta\big) + P\big(|\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s| > \eta\big) \leq \\ C\eta + \eta^{-2}\|\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s\|_{2,P}^2\end{aligned} \tag{8.4}$$

where we used Chebyshev's inequality to get the last line. Consequently

$$I_1 \leq \sum_{s=t+1}^{T-1} \mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}}(X)_s - q_{\mathcal{H}}(X)_s\|_{2,P}^2 > \frac{\eta^2\varepsilon}{\|f\|_{\infty,X}^2} - C\eta^3\big)$$

Optimizing over $\eta$, yields the optimal value of $\eta = \frac{2\varepsilon}{3C\|f\|_{\infty,X}}$. The bound for $I_2$ follows directly from (6.17). □

**Theorem 8.3.** *Let $f \in L_\infty(X)$ and assume that hypothesis 8.1 holds. If $k_n, H_n \to \infty$ and*

$$\frac{k_n H_n^4 \log(H_n)}{n} \to 0 \tag{8.5}$$

*then $\|\hat{q}_{n,\mathcal{H}^{(n)}}(X)_t - q(X)_t\|_{2,P}^2 \xrightarrow{\mathbb{P}} 0$ for $n \to \infty$.*

*Proof.* To derive our asymptotic result, (5.31) is preferred because the constants are explicit. The reduction in the exponential decay factor is irrelevant for the purpose of our limit result. The class $\mathcal{L}_t^{(n)}$ is uniformly bounded by $L_n = (H_n + \|f\|_{\infty,P})^2$. Bound the covering numbers

$$N(\varepsilon L_n, \mathcal{L}_t^{(n)}, d_{1,P_n}) \le N\big(\frac{\varepsilon(H_n + \|f\|_{\infty,P})}{2}, \mathcal{H}_t^{(n)} - \mathcal{Y}_t^{(n)}, d_{2,P_n}\big) \le \\ N\big(\frac{\varepsilon H_n}{4}, \mathcal{H}_t^{(n)}, d_{2,P_n}\big) N\big(\frac{\varepsilon \|f\|_{\infty,P}}{4}, \mathcal{Y}_t^{(n)}, d_{2,P_n}\big) \le \\ e^2(k_n+1)(k_nc+1)\left(\frac{8\sqrt{2e}}{\varepsilon}\right)^{2k_n(1+c)}.$$

and apply (5.31)

$$\mathbb{P}\big(\sup_{l\in\mathcal{L}_t^{(n)}}|P_nl - Pl| > \frac{\varepsilon}{2}\big) \le 8\mathbb{E}[N(\frac{\varepsilon}{16}, \mathcal{L}_t^{(n)}, d_{1,P_n})]\ \exp\big(-\frac{n\varepsilon^2}{512L_n^2}\big) \le \\ e^2(k_n+1)(k_nc+1)\left(\frac{128\sqrt{2e}L_n}{\varepsilon}\right)^{2k_n(1+c)} \exp\big(-\frac{n\varepsilon^2}{512L_n^2}\big).$$

where $c = 2(T-t)\log_2(e(T-t))$. Proceed as in in the proof of theorem 6.15 to conclude that

$$\mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}^{(n)}}(X)_t - q_{\mathcal{H}^{(n)}}(X)_t\|_{2,P}^2 > \varepsilon\big) \le \\ (T-t)2^{T-t}e^2(k_n+1)(k_nc+1)\left(\frac{128\sqrt{2e}L_n}{\varepsilon}\right)^{2k_n(1+c)} \exp\big(-\frac{n}{512L_n^2}\,\varepsilon_{T-1}^2\big)$$

where $\varepsilon_s$ satisfies the recursion $\varepsilon_t = \varepsilon$ and $\varepsilon_{s+1} = \frac{4\varepsilon_s^3}{27C^2\|f\|_{\infty,X}^6}$. Inspecting the right hand side shows that

$$\mathbb{P}\big(\|\hat{q}_{n,\mathcal{H}^{(n)}}(X)_t - q_{\mathcal{H}^{(n)}}(X)_t\|_{2,P}^2 > \varepsilon\big)$$

converges to zero if (8.5) holds. The convergence of the approximation error to zero is clear. □

## 9. Outlook

Instead of using global linear approximation architectures, the algorithm could be based on nonlinear architectures such as neural networks, radial basis functions or $n$-term approximates. However solving the minimization problems (3.5) becomes much harder because of the existence of many local extrema.

Another direction of generalization are local approximation schemes, such as $k$-nearest neighbor and other kernel estimators like the Nadaraya-Watson estimator or more generally local polynomial kernel regression estimators, [15]. These types of architectures are appealing because their asymptotic optimality. Stone showed in [38] that if the regression function $m(x) = E[Y \mid X = x]$ is $p$-smooth, $X, Y$ are bounded and $X$ has a density, then the $L_2$-error of a local polynomial kernel estimator converge to zero at a rate of $n^{2p/(2p+d)}$, and that this rate is optimal in a minimax sense.

We would like to stress that both generalizations can be approached in the framework of structural risk minimization.

Another improvement, which can be combined with the above extensions, is obtained by replacing the stopping times $\tau_{t,h}$ of (2.18) with the *fuzzy* stopping times

$$\sigma_{t,\alpha,h}(x) = t\theta_\alpha(f(x)_t - h(x)_t) + \sigma_{t+1,\alpha,h}(x)\big(1 - \theta_\alpha(f(x)_t - h(x)_t)\big) \qquad (9.1)$$

where $\theta_\alpha(s) = (1+\exp(-\alpha s))^{-1}$ and $\alpha$ is a parameter. If $\alpha \to \infty$, then $\theta_\alpha$ converges to the step function at 0, uniformly on the complement of every neighborhood of 0. This smoothing approach is very similar to neural network models, where the hard threshold function of the McCulloch–Pitts model is replaced by the sigmoidal activation function. The advantage of fuzzy stopping times is that the estimates (2.22), (2.23) can be significantly improved. This leads to sharper exponential inequalities for the sample error.

We plan to explore these extensions and refinements in a subsequent article.

## References

1. M. A. Arcones, *A remark on approximate m-estimators*, Statistics and Probability Letters (1998), no. 38, 311–321.
2. G. Barnone-Adesi and R. Whaley, *Efficient analytic approximation of American option values*, J. of Finance **42** (1987), 301–320.
3. A. Bensoussan and J.L. Lion, *Application of variational inequalities in stochastic control*, Studies in Math. and its Appl., vol. 12, North-Holland, 1982.
4. A. Benveniste, M. Metiver, and P. Priouret, *Adaptive algorithms and stochastic approximations*, Sringer-Verlag, Berlin, 1990.
5. P. Boessarts, *Simulation estimators of optimal early exercise*, Working paper, Carnegie-Mellon University, 1989.
6. P. Boyle, *Options: a Monte Carlo approach*, Journal of Financial Economics **4** (1977), 323–338.
7. M. Broadie and P. Glasserman, *Pricing American-style securities using simulation*, J. of Economic Dynamics and Control **21** (1997), 1323–1352.
8. ———, *Monte Carlo methods for pricing high-dimensional American options: An overview*, Monte Carlo Methodologies and Applications for Pricing and Riks Management, Risk Books, 1998, pp. 149–161.
9. J. Cox, S. Ross, and M. Rubinstein, *Option pricing: A simplified approach*, J. Financial Economics **7** (1979), 229–263.
10. M. Dempster and J. Hutton, *Pricing American stock options by linear programming*, 1999, pp. 229–254.
11. L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, Stochastic Modelling and Appl. Probability, vol. 31, Springer, 1996.
12. R. M. Dudley, *Central limit theorems for empirical measures*, Annals of Prob. **6** (1978), no. 6, 899–929.
13. ———, *Uniform central limit theorems*, Cambridge Studies in advanced mathematics, vol. 63, Cambridge Univ. Press, 1999.
14. Panchenko D., *Symmetrization approach to concentration inequalities for empirical processes*, Preprint (2002), 1–15.
15. J. Fan and I. Gijbels, *Local polynomial modelling and its applications*, Chapman & Hall, 1996.
16. G. Freud, *On polynomial approximation with respect to general weights*, Functional analysis and its applications international conference, Madras, 1973 (Garnir H. G. et. al., ed.), Lecture notes in mathematics, vol. 399, Springer, 1974, pp. 149–179.
17. Rogers L. C. G., *Monte Carlo valuing of American options*, University of Bath, Preprint, 2001.
18. D. Haussler, *Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension*, Journal of Combinatorial Theory **69** (1995), no. 2, 217–232.
19. J. Huang and J. Pang, *Option pricing and linear complementarity*, 1998.
20. P. Jaillet, D. Lamberton, and B. Lapeyre, *Variational inequalities and the pricing of American options*, Acta Appl. Math. **21** (1990), 263–289.
21. I. Karatzas and S. E. Shreve, *Methods of mathematical finance*, Applications of Math., vol. 39, Springer-Verlag, 1998.
22. I. Karatzas, *On the pricing of American options*, Applied Mathematics and Optimization **17** (1988), 37–60.
23. H. J. Kushner and D. S. Clark, *Stochastic approximation methods for constrained and unconstrained systems*, Springer-Verlag, New York, 1978.
24. H. J. Kushner and G. G. Yin, *Stochastic approximation algorithms and applications*, Stoch. Modelling and Appl. in Prob., vol. 35, Springer, 1997.
25. H. J. Kushner, *Numerical methods for stochastic control in finance*, Mathematics of Derivative Securities, Cambridge Univ. Press, 1997, pp. 504–527.

26. D. Lamberton and G. Pagès, *Sur l'approximation des réduites*, Ann. Inst. Henri Poincarés, Probab. Statist. **26** (1990), no. 2, 331–355.
27. S. B. Laprise, Y. Su, R. Wu, M. C. Fu, and D. B. Madan, *Pricing American options: A comparision of Monte Carlo simulation approaches*, Journal Comput. Finance **4** (2001), no. 3, 39–88.
28. F. A. Longstaff and E. S. Schwartz, *Valuing American options by simulation: A simple least-square approach*, Review of Financial Studies **14** (2001), no. 1, 113–147.
29. H. N. Mhaskar and Pai D. V., *Fundamentals of approximation theory*, CRC Press, 2000.
30. H. N. Mhaskar, *Weighted polynomial approximation*, Journal Approx. Theory **46** (1986), 100–110.
31. ———, *Introduction to the theory of weighted polynomial approximation*, Series in Approximations and Decompositions, vol. 7, World Scientific, 1996.
32. ———, *Private communication*, 2002.
33. S. Mulinacci and M. Pratelli, *Functional convergence of the Snell envelope: Applications to American options approximations*, Finance and Stochastics (1998), no. 2, 311–327.
34. R. Myneni, *The pricing of the American option*, Annals of Appl. Probability **2** (1992), no. 1, 1–23.
35. D. Pollard, *Convergence of stochastic processes*, Springer Series in Statistics, Springer-Verlag, 1984.
36. ———, *Empirical processes: Theory and applications*, NSF-CBMS Regional Conference Series in Statistics, Inst. of Maths. and Am. Stat. Assoc., 1990.
37. P. Protter, D. Lamberton, and E. Clément, *An analysis of the Longstaff-Schwartz algorithm for American option pricing*, Finance and Stochastics **6** (2002), no. 4, 449–471.
38. C. J. Stone, *Optimal rates of convergence for nonparametric regression*, Annals of Statist. **10** (1982), no. 4, 1040–1053.
39. M. Talagrand, *Sharper bounds for Gaussian and empirical processes*, Annals of Prob. **22** (1994), no. 1, 28–76.
40. J. A. Tilley, *Valuing American options in a path simulation model*, Transactions of the Society of Actuaries **45** (1993), 83–104.
41. J. N. Tsitsiklis and B. Van Roy, *Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives*, IEEE Trans Autom. Control **44** (1999), no. 10, 1840–1851.
42. A. W. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes with applications to statistics*, Springer Series in Statistics, Springer, 1996.
43. P. L. J. Van Moerbeke, *On optimal stopping and free boundary problems*, Archive Rat. Mech. Anal. **60** (1976), 101–148.
44. V. N. Vapnik and Chervonenkis A. Y., *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability and its Applications **16** (1971), no. 2, 264–280.
45. V. N. Vapnik, *Estimation of dependences based on empirical data*, Springer Series in Statistics, Springer-Verlag, 1982.
46. ———, *The nature of statistical learning theory*, 2nd ed., Statistics for Engineering and Information Science, Springer, 2000.

ZKB, P.O. Box, CH-8010 Zürich, Switzerland
*Current address*: Kehlhof 2a, 8409 Winterthur, Switzerland
*E-mail address*: daniel.egloff@dplanet.ch

Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario L8S 4K1, Canada
*E-mail address*: minoo@math.mcmaster.ca